PAUL R. VOSS
KATHERINE J. CURTIS WHITE
ROGER B. HAMMER

EXPLORATIONS IN SPATIAL DEMOGRAPHY

INTRODUCTION

Social scientists in many disciplines have noted a re-emerging interest in issues concerning social processes embedded within a spatial context (e.g., Messner & Anselin, 2004). In this chapter, we echo and emphasize the long-standing assertion, found in various forms across numerous disciplines, that special methods are necessary for the appropriate analysis of spatial data. Attributes of spatially referenced data generally violate at least one of the assumptions underlying the standard regression model, which necessitates both caution regarding these violations and attention to methods designed to correct for them. We discuss the nature of the problem, how it arises, how to identify it, and methods by which one can press forward appropriately with the investigation of such data. We present what we view as the most important and well-developed concepts of spatial data analysis and indicate for interested readers where greater detail can be found. Specifically, we have sought to minimize the presentation of technical material, including formulae and equations, and, instead, apply the concepts and methods to an analysis of population change in the Great Plains.

SPATIAL IS SPECIAL

When investigating population change for a large number of spatial units (e.g., counties), it is the natural inclination of sociologists and demographers to move from simple descriptive analyses to begin asking such questions as: How might these data be modeled? How well can I account for variability in attribute values among geographic units by identifying other covariates of our attribute of interest? Such analysts have traditionally turned to multivariate regression modeling to answer such questions. Regrettably, standard regression approaches to data for spatial units bring special complications that have not always been appreciated or understood.

The idea that somehow “spatial is special” is a notion that has begun only slowly to enter the awareness of quantitative demographers.

Over the past two decades, increasing attention has been drawn to the fact that spatial data require special analytical approaches. Many of the techniques documented in standard statistics textbooks and taught in our “methods” classrooms unfortunately confront significant difficulties when applied to the analysis of geospatial data. These problems are summarized by language more familiar to geographers and regional scientists than to demographers: spatial autocorrelation, the modifiable areal unit problem, scale and edge effects. But the emphasis on “problems” fails to capture the fact that there also is a benefit arising from the special nature of spatial data. Aspects of space (e.g., distance, proximity, and interaction), when properly acknowledged and incorporated into one’s model, can overcome the complications of space and error dependence, improve the specification of models based on spatial units, and provide estimates of parameters that are less subject to statistical bias, inconsistency, or inefficiency. Further, such approaches can contribute to theoretical notions regarding the role of space in social relationships and processes.

Although rural demography has long maintained a strong focus on patterns and trends that vary spatially (Voss, 1993; 2004), the field has not been very sensitive to these more recent analytical issues, and rural demographers have largely failed to adopt the methods of formal quantitative spatial analysis that have emerged in the fields of geography, regional science, and spatial econometrics during the past decade or so (Lobao & Saenz, 2002). It is encouraging that such neglect is waning, as evidenced by the spatial focus of a recent Rural Sociological Society presidential address (Lobao, 2004).

To illustrate some of these spatial concepts, we examine in this chapter the correlates of county-level population change in the Great Plains between 1990 and 2000. Details regarding the sample, measures, and theoretical motivations can be found in White (2003).

A thorough researcher will carefully begin such an analysis by exploring the behavior of the variables of interest using the standard tools of exploratory data analysis (EDA) – and thus we begin. In the present example, one that will be used throughout the remainder of the chapter, interest is focused on population change (measured as the natural log of P_{2000}-P_{1990}/ P_{1990}) and a few potentially useful, theoretically derived covariates of population change: farm dependence, population age structure, climatological conditions, metropolitan status, county acreage (natural log) and initial county population (natural log). The latter two variables are of less substantive interest and are included in the model as possible controls for heteroskedasticity.

When undertaking initial EDA explorations of spatial data, in addition to examining the univariate statistical distributions of the attributes (for normality, outliers, etc.) and their bivariate relationships with the dependent variable (for linearity), it also is worthwhile to develop a sense of the spatial distributions of the attribute values. As illustrated in Figure 1 (the county boundaries used throughout this example refer to 1900 boundaries since the example is taken from a larger, historic project), the map of population change indicates that roughly one-in-twelve Great Plains counties suffered population loss in excess of 10 percent over the decade of the 1990s, while more than one-in-four counties witnessed population growth of more than 10 percent during the same period. Growth characterizes many of the east-west boundary counties, while loss is largely concentrated along a north-south axis among the central counties and along the northern edge of the region. These concentrations lead to two initial conclusions: First, there is sub-regional variation within the larger Great Plains region (something we discuss below as *spatial heterogeneity*). Second, there appears to be evidence of spatial clustering, such that counties experiencing growth seem to be near other counties experiencing growth while those suffering loss are near other counties undergoing loss (which we discuss below as possible evidence of *spatial dependence*). By mapping our data and reviewing the distributions of the variables across space, it becomes evident that spatial patterning (in the form of spatial autocorrelation) will have to be addressed in our modeling strategy.

[Figure 1 About Here]

Spatial Autocorrelation

Those who have studied time-series analysis will recognize the parallels to temporal autocorrelation. Typically, when most social phenomena are mapped, locational proximity usually results in value similarity. High values tend to be located near other high values, while low values tend to be located near other low values, thus exhibiting *positive* spatial autocorrelation (Cliff & Ord, 1973; 1981). Such appears to be the case with population change in the 1990s within the Great Plains. Less often, high values may tend to be co-located with low values (or vice versa) as “islands” of dissimilarity or in a spatial “checkerboard” pattern that exhibits *negative* spatial autocorrelation (see Tolnay, Deane & Beck, 1996). In either case, the units of analysis in spatial demography likely fail a formal statistical test of randomness and thus fail to meet a key assumption of classical statistics: independence among observations. With respect to statistical analyses that presume such independence (e.g., standard regression analysis), positive autocorrelation means that the spatially autocorrelated observations bring less information to the model estimation process than would the same number of independent observations. The greater the extent of spatial autocorrelation, the more severe is the information loss. Again, this fact has been known for several decades. For example, early recognition of this problem is found in a brief paper by census statistician Frederick Stephan, who, when referring to the use of census tract data in social research, introduced the problem by analogy to classical sampling theory: “Data of geographic units are tied together, like bunches of grapes, not separate, like balls in an urn” (1934, p. 165).

How Does Spatial Autocorrelation Arise?

We have pointed out that *positive* spatial autocorrelation is very commonly a property of mapped social and economic data, whereas negative spatial autocorrelation is much less commonly observed. A quick explanation for the presence of spatial autocorrelation can be found in the oft-cited “first law of geography,” enunciated by Tobler in 1970: Everything is related to everything else, but near things are more related than distant things (1970, p. 236). While useful as a short-hand reminder, Tobler’s first law is somewhat unsatisfying because it doesn’t tell us *why* this phenomenon arises in practice, or what difference it makes. Why, for example, do state sales tax levels tend to cluster regionally? Why does the percentage vote cast for presidential candidates show systematic geographic clustering? Why do high housing values cluster in some neighborhoods of a large city and low values in other neighborhoods? Or, as in the case of our example, why is relatively high growth concentrated in some sub-regions of the Great Plains and low growth (or decline) in others?

While there exist some helpful reviews on this topic (e.g., Wrigley, Holt, Steel, & Tranmer, 1996, pp. 30-31; Brueckner, 2003), the answers to such questions can only be approximated with models of the spatial process that inevitably are imperfect. Such answers generally will be a function not only of the data being analyzed but will depend strongly on the analyst’s theory about the process, as well as assumptions underlying both the data and the statistical model(s) selected to describe the nature of the relationships under investigation. For example, the four substantively interesting independent variables selected for our example (farm dependence, population age structure, climatological conditions, metropolitan status) and two additional control variables were not chosen at random but have been identified in earlier work addressing population change. Our task is to analyze appropriately the nature of their joint relationship with population change while simultaneously accounting (or correcting) for spatial process relationships at work in the data.

Exploratory Spatial Data Analysis

While much of the growing literature on spatial data analysis focuses on matters of specification tests, parameter estimation, and advanced tools such as Monte Carlo simulation, any proper empirical analysis must begin more simply by exploring and understanding one's data. Continuing our earlier discussion of EDA, many of the techniques first codified by John Tukey (1977) and later expanded by Tukey's colleagues (Hoaglin, Mosteller & Tukey, 1983; 1985) are also appropriate for the exploration of spatial data. Once again, however, some of the unique aspects of spatial data make exploratory *spatial* data analysis (ESDA) a field that has attracted considerable attention in and of itself. The science of creating and interpreting maps of spatial data, for example, is the topic of a large literature fostered by the development over the past 30 years of powerful geographic information systems (GIS) (Chou, 1997). In addition, software for creating and testing a variety of neighborhood weights matrices, for generating various measures of spatial autocorrelation (both global and local), and for obtaining diagnostic results concerning error dependence in standard regression models are now widely available. This literature is large and dynamic. Perhaps the best citation that can be provided is to invite the reader's attention to the website of the Center for Spatially Integrated Social Science (CSISS), a center whose mission is to serve as an ongoing clearinghouse for software tools, literature, and training opportunities in spatial data analysis (http://www.csiss.org).

Global and Local Diagnostics

Global measurements – whether they are overall descriptions of attribute values, measures of statistical relationships, or model accuracy assessments – are derived using data for the entire study region. For example, a global Moran's *I* statistic is a single measure describing the general extent of spatial clustering of an attribute across the region, conditional on the specific neighborhood structure imbedded in the chosen weights matrix (Moran, 1950). The global Moran's *I* can be scaled to the interval (-1,1) where a strong positive value indicates value similarity among neighbors (clustering, or *positive* spatial autocorrelation), a strong negative value indicates value dissimilarity (dispersion, or *negative* spatial autocorrelation), and a value near zero suggests no spatial relationship. Tests for significance use z-scores and the standard normal distribution. As commonly applied to a full data set, Moran's *I* yields an indication of the extent of overall spatial clustering of similar values on a given attribute. It is a "global" measure of spatial autocorrelation and, as such, cannot by itself identify *where* "hot spots" of value clustering exist within the study region. Since spatial data are easily mapped, it is thus only natural that techniques have been developed for generating and mapping *local* counterparts to many global measurements.

Two useful ESDA tools in spatial data analysis are the Moran Scatterplot (Aneslin, 1996) and so-called LISA statistics (for Local Indicators of Spatial Association) such as the "local" Moran's *I* (Anselin, 1995). These devices are extremely valuable for gaining an understanding of the localized extent and nature of spatial clustering in a data set. Their use logically should precede and inform the process of hypothesis construction, model specification, estimation, and statistical inference. Rather than producing a single global statistic or parameter, local analysis generates statistics or parameters that correspond with researcher-specified smaller-scale local areas (commonly called "neighborhoods"). It is helpful to re-emphasize that it is the researcher, not the data or some accommodating software program, who defines what is meant by a local neighborhood. As indicated earlier, this is done by specifying a matrix of weights (≤ 1) that characterizes the structure of local dependence. There exists a large literature on the topic of selecting a weights matrix, and Griffith (1996) is but one helpful resource.

Figure 2 shows the Moran scatterplot for the Great Plains dependent variable: log percent growth (for counties) 1990 to 2000. In this exploratory view, the data are standardized so that units on the graph are expressed in standard deviations from the mean. The horizontal axis shows the standardized value of the log percent population change for each county. The vertical axis shows the standardized value of the *average* log percent population change for that county's "neighbors" as defined by the weights matrix. Neighbors for this illustration are defined under the "first-order queen" convention, meaning that the neighbors for any given county "A" are those other counties that share a common boundary (or single point of contact) with "A" in any direction. Importantly, "A" is not considered a neighbor of itself and is excluded from the average. Counties on the border of the Great Plains region, as shown in Figure 1, are permitted only to have neighbors within the region. This restriction creates some boundary problems ("edge effects") in this analysis, but the topic is not addressed further in this overview. The reader is referred to any of several articles or texts on spatial data analysis for further information and ways of dealing with such problems (e.g., Martin, 1987).

The upper right quadrant of the Moran scatterplot shows those counties with above average growth which share boundaries with neighboring counties that also have above average growth (high-high). The lower left quadrant shows counties with below average growth and neighbors with below average growth (low-low). The lower right quadrant has counties with above average growth surrounded by counties with below average growth (high-low), and the upper right quadrant has the reverse (low-high). Anselin (1996) has demonstrated that the slope of the regression line through these points conveniently expresses the global Moran's *I* value, which, for our Great Plains example, is 0.54. This statistic is strongly positive, indicating powerful *positive* spatial autocorrelation (clustering of like values). Most counties are found in the high-high or low-low quadrants.

[Figure 2 About Here]

In Figure 3 we show a LISA cluster map which displays in a different way the same data as the Moran scatterplot of Figure 2. The map shows where in the Great Plains region the various combinations of high-high, low-high, etc. counties are found. Counties where the local Moran statistic is not significant (at the .05 level, based on a randomization procedure) are not shaded. Hotspot clusters of high growth counties surrounded by high growth counties are apparent in the sprawling east-central Texas region connecting metropolitan areas of Dallas-Fort Worth, Austin, San Antonio and Houston-Galveston. Another large high-high cluster connects the Denver-Boulder, Colorado Springs and Pueblo metropolitan areas, and a small high-high cluster is found mostly in the Missouri counties southeast of Kansas City. Coldspots include the (low-low) clusters of counties of central North and South Dakota, central Nebraska and Kansas, and two or three small clusters in the Texas Panhandle region and other areas of east-central Texas.

[Figure 3 About Here]

Individual high-low counties appear as islands throughout a central band running north-south through the Great Plains. Often these include small, somewhat isolated, metropolitan counties – e.g., Burleigh County (Bismarck) and Cass County (Fargo) in North Dakota. A few statistically significant (at the .05 level) low-high counties are also present in the region. These defy easy summarization, save for the fact that they are largely found along or near the borders of the region (and thus may suffer from unknown but troublesome edge effects). While this exploratory view of the data may suggest hypotheses for the analyst to confirm in the inferential part of any further analysis, perhaps the principal message for us at this point is that, taken together, the maps in Figure 1 and Figure 3 confirm that the process of growth in the Great Plains in the 1990s has conspired somehow to partition the region into identifiable sub-regions of growth and decline. Such spatial heterogeneity must be addressed in any further analysis of the data, and we begin by examining whether there might be parameter regimes that can be associated with the patterns observed in Figures 1 and 3.

Geographically Weighted Regression

One of the more recent and fascinating developments in the design of local statistics is the theoretical/conceptual background, and associated software, to explore how regression parameters and regression model performance vary across a study region. Geographically Weighted Regression (GWR) is similar to a global regression model in that the familiar constant, regression coefficients, and error term are all present within the regression specification. In addition to classical linear regression, GWR is also related to spatial regression such that the spatial weights are 0 in the former while the latter is equivalent in the solution for localized slopes via the spatial multiplier $(I – \rho W)^{-1}\ \beta_k$ (Deane, Beck & Tolnay, 1998). There are two ways in which GWR differs from standard (global) regression, however. First is the fact that a separate regression is carried out at each location (observation) using only the other observations that lie within a user-specified distance from that location. Second, the regression specification includes a statistical device which weights the attributes of nearby counties more highly than it does the attributes of distant counties. The result is a set of *local* regression parameters for each county. The precise implementation of GWR is controllable by the analyst and is far too detailed for discussion here (see Fotheringham, Brunsdon & Charlton, 2002). The important feature to emphasize, however, is that the output file enables the researcher to examine and map local parameter estimates and local regression diagnostics, thereby enabling assessment of the utility of the model for various portions of the larger study region.

Examples of such maps are illustrated in Figures 4 through 7. Local R^2 statistics are mapped across the region in Figure 4, illustrating those areas where the model performs well versus those where the model "fit" is less precise. The local R^2 statistic in this example ranges widely from 0.230 to 0.740. We note that the model's highest performance is found roughly in southern Oklahoma and in the northwestern Plains counties in western North Dakota and eastern Montana. Lower model fits are generally found among the boundary counties but specifically in the Texas Panhandle region, in southern Iowa, and, to a lesser extent, in western Nebraska. When referring back to the distribution of population growth (Figure 1), variation in model fit does not appear to associate closely with either areas of growth or areas of loss. For instance, the model fits relatively poorly (low R^2) both in the loss (Panhandle) and the growth (southeastern) clusters of Texas counties.

[Figures 4 through 7 About Here]

GWR parameter estimates can also be mapped and compared to gain further insight regarding spatial variation in relationships. We stress that these tools are exploratory in nature as opposed to explanatory. GWR can be a useful guide in showing where particular covariates of the response variable contribute strongly and where they do not. The parameters shown in Figures 5 through 7 are the intercept term and those for two of the independent variables, farm employment and temperature range, respectively. Caution is advised when attributing statistical significance to the local parameter values because of the high degree of multiple hypothesis testing in GWR. Some type of Bonferroni-like adjustment to the critical values clearly is appropriate. Fotheringham and colleagues suggest rejecting the null only when t-values approach 4.5 and greater (2002:135).

The map showing the distribution of the intercept parameter (Figure 5) indicates that, controlling for the response to predictive variation from the six independent variables, the level of the local intercept varies rather dramatically across the Great Plains (from negative .956 to positive 2.158). Such intercept heterogeneity suggests the likely presence of an unaccounted interaction in the model. For example, local intercept values are relatively high for the band of counties sweeping toward the northeast from southern Texas to northwestern Missouri. The intercept also is high in the higher growth area around (and north of) the Denver metropolitan area. Among these counties, the local parameters for a number of our variables are negative in value and moderately strong (for example, see Figure 6, which shows local variation of the parameter for the farm employment variable). On the other hand, local intercept values are relatively low (and negative) in northern Texas, southwestern Kansas, and southern Minnesota. One variable appears to be contributing strongly to these lower local intercepts: the temperature range variable. For this predictor variable, the response of regional growth is strong and positive (Figure 7).

While visualizing a regression hyperplane in seven dimensions is challenging, to say the least, talking about it in general terms may be easier. An examination of the maps of the GWR-generated local parameters (of which only three are presented in Figures 5 through 7) suggests the following types of local interactions. In the areas of northern Texas, southwestern Kansas and southern Minnesota, our Ordinary Least Squares (OLS) regression hyperplane has a positive slope (especially strong, marginally, on the temperature range dimension). The positive slope produces a negative intercept value in these portions of the region. On the other hand, in those portions of the region where the intercept is positive and relatively strong (southern Texas to northwestern Missouri and in the vicinity of the Denver metropolitan area), the hyperplane likely has an overall negative slope. These implied interactions might well inform a respecification of our model to accommodate the interactions. While this is a promising direction, we do not embark on this particular path in the remaining analysis reported here. Rather, we seek to deal directly with the implied spatial heterogeneity by fitting a trend surface to our data before tackling any spatial dependence that may remain after modeling the spatial heterogeneity.

Spatial Heterogeneity versus Spatial Dependence

As hinted at in the preceding section, large-scale regional differentiation (among attribute values and/or among parameter values) is an important component of spatial variation. Most treatments of spatial data analysis refer to such sub-regional variation as “spatial heterogeneity.” We follow the usual convention of referring to spatial heterogeneity as the lack of stability across space of one or more attribute values (more formally expressed as lack of stability in the moments of the joint probability distribution of the attributes) or as lack of stability of relationships among the attributes as measured by correlation statistics or regression parameter values (see Anselin, 1988). Spatial heterogeneity often is a concept referred to somewhat casually or vaguely – as we are guilty of here. A more precise sense of what is captured in the notion of spatial heterogeneity is contained in the statistical concept of spatial stationarity in its various forms (Cressie, 1993). In essence, the term "heterogeneity" simply gives recognition to the common observation that values of a variable, or values of relationships among variables, are not the same across space. Few social processes are spatially *homogeneous*.

In our example, the nature and extent of population change and its associations with correlated factors are distributed unequally across the Great Plains. In particular, the term spatial heterogeneity applies to large-scale trend or drift in a spatial process, where "large-scale" is taken to mean scales involving distances that extend well beyond any "neighborhood" structure imposed on the data (as discussed further below). Spatial heterogeneity often is also referred to as "first-order variation" or as "first-order spatial effects" in a spatial process (Bailey & Gatrell, 1995). The inclusion in a regression model of one or more variables might satisfactorily account for the observed spatial heterogeneity. If population growth is mainly concentrated in specific types of counties, for example, and if this is the spatial process dominating our data, then inclusion of a dummy variable to identify these counties would not only boost the explanatory value of the model but also would reduce the extent of spatial heterogeneity and, ideally, also reduce or eliminate heteroskedasticity and spatial autocorrelation among the residuals. Another approach to deal with large-scale trends is to fit a trend surface to the data, as we illustrate below.

"Spatial dependence" ("second-order variation") refers to small-scale spatial effects that manifest as a lack of independence among observations. The assumption is that dependence among the observations derives from spatial interaction among the units of analysis which ideally can be defended theoretically and which can be statistically captured by a spatially lagged "neighborhood" effect in a model of the spatial process. Such spatial lags may involve the dependent variable, one or more of the independent variables, the error term, or some combination of all three. Properly specified and estimated, such a model with spatial lags is able to “borrow information” or “borrow strength” from neighboring observations precisely because of the spatial autocorrelation among the units of analysis (Haining, 2003, p. 36). We do not present the details, but once a spatial lag is included in a regression model to account for spatial dependence in the data, maximum likelihood estimation (MLE) is usually the appropriate estimator (see Anselin & Bera, 1998). In our example, a carefully selected variable to account for spatial heterogeneity in the data might boost the explanatory value of the model and largely remove the large-scale spatial process, but spatial autocorrelation would persist if a spatial dependence process also were indicated. In other words, there would remain in the data a more complicated, interactive spatial relationship among neighbors that suggests the requirement of some type of autoregressive term in the regression specification.

While the preceding discussion appears to present a sequential, orderly, step-by-step process, in practice the situation is more complex. Often the data suggest a combination of both first-order and second-order effects or fail to give unambiguous clues to one or the other. For example, the map of recent population change in the Great Plains (Figure 1) reveals an uneven gradation of population growth and decline in the 1990s that defies any simple and immediate explanation. Several clusters of counties with high growth are apparent: e.g., east-central Texas, central Missouri, eastern Colorado – certainly very different counties in terms of topography, cultural history, and industrial base. Clusters of slow growth or population decline are apparent across the most northern Plains counties (Montana, North Dakota, and northwestern Minnesota), in much of Nebraska and Kansas, and in the Texas Panhandle. Might these clusters be accounted for by established historical or legacy effects, and might they be “explained” by a few well chosen independent variables? Or might there exist neighbor influences among these counties (e.g., spatial spillovers or diffusion) that account for the spatial pattern? The first question inquires about possible spatial heterogeneity, the second about possible spatial dependence. For whatever reasons, some parts of the Great Plains reveal growth (some with relatively high growth), and other parts show low growth or population decline. The goal of the researcher is to identify potential covariates of population change in the region and to explain the variation in growth among Great Plains counties using a combination of traditional modeling approaches and newer spatial modeling approaches.

Regardless of the analyst's theoretical notions about the process giving rise to the observed spatial pattern, the analysis generally proceeds as follows. First, based on a combination of theory and review of the relevant literature, a defensible OLS regression model is fit to the data, and a variety of residual-based diagnostics are examined, including a test for spatial autocorrelation. Tests for spatial error dependence generally take two forms: (1) a general test for spatial autocorrelation of residuals against the alternative of no autocorrelation, and (2) a set of tests against a specific *form* of spatial process. The first such generalized test usually is the calculation of a region-wide or "global" measure of spatial autocorrelation, such as the Moran *I* statistic, as discussed above. The second set of specific tests is based on the maximum likelihood principle (see Anselin, 2001; Anselin & Bera, 1998). We comment on these tests in interpreting the regression model results below.

Unfortunately, in the cross sectional context, there do not exist statistical tools to inform the analyst which spatial process, heterogeneity or dependence, has generated the data at hand (Bailey & Gatrell, 1995, p.32-33). That is, it is not mathematically possible to differentiate an independent heterogeneous spatial process from a dependent homogeneous spatial process. As mapped realizations, they may appear quite identical. Either process alone (or both acting together) could be responsible for the spatial pattern shown in Figure 1. The story is less bleak if repeated observations (over time) are available for cross sectional data. There may, under such conditions, be sufficient data to distinguish between the two spatial processes. Moreover, the distinction between large-scale variation and small-scale variation in an attribute is rarely easily determined. It depends in part on how the analyst has chosen to define "neighborhood" structure. As described earlier, the latter is expressed formally in a proximity or weights matrix. This matrix captures the researcher's view of the *nature* of neighboring influences. The actual *degree* of such influences is captured by the data and a spatial parameter to be estimated along with other parameters. A strong theoretical framework and some testing of alternatives should guide the choice of spatial weights, as they play a strong role in determining statistics or parameter values derived using a specific weights matrix. This matrix is required for the calculation of spatial autocorrelation statistics, such as Moran's *I*, and for specifying and estimating regression models incorporating spatial dependence terms to account for spatial autocorrelation in the data.

Thus far in our discussion, spatial autocorrelation has been described as something that arises from a substantive spatial process. In the case of spatial heterogeneity, there are presumed forces (geophysical, cultural, social, or economic) that somehow work to constrain or otherwise serve as influences causing individuals (or families or counties) with similar attribute bundles to find themselves (by choice or otherwise) to be physically located near one another. In the case of spatial dependence, presumed *interaction* among individuals results in spatial clustering. The large body of literature springing from the theory of social adoption/diffusion (Rogers, 1962), for example, captures well the notion of spatial dependence.

However, spatial autocorrelation can also arise as a nuisance (Anselin, 1988). Most commonly this occurs when the underlying spatial process creates regions of value clustering that are much larger than the units of observation chosen by (or available to) the analyst. An example of such nuisance autocorrelation might be present in the distribution of population growth in the Great Plains. The large cluster of high growth counties in central Texas (Figure 1) is discussed above as a sub-region contributing to spatial heterogeneity, and this sub-region contributes heavily to the fairly high global Moran's *I* statistic. Stepping back from the data for a moment, one quickly observes that this sub-region of high growth is considerably more extensive than is the particular lens (counties) through which the process is viewed. When units of analysis are smaller than the boundaries of areas having high or low attribute values, spatial autocorrelation in the observations is inevitable. Such nuisance autocorrelation must somehow be recognized and eventually brought into the formal analysis of the data. Customarily this is handled in models of spatial heterogeneity with the use of dummy variables to identify different "spatial regimes" or through the incorporation of a "surface trend" as part of the regression model (Anselin, 1988).

The aim of the researcher is to specify and estimate a model that reasonably accounts for or incorporates the spatial effects present in the data. These effects can be modeled separately or jointly as spatial heterogeneity and spatial dependence. When first examining a spatial relationship, the researcher must ask whether the association appears to be a *reaction*, characteristic of heterogeneity, or an *interaction*, indicative of spatial dependence. Anselin, referring to earlier studies, discusses this difference using the terms "apparent contagion" (spatial heterogeneity) and "real contagion" (spatial dependence) (1988, p. 15).

If the association is merely a reaction to some geophysical, cultural, social, or economic force that works to create spatial patterning, then a modeling strategy with a standard regression structure may be appropriate. Often it is discovered that independent variables in the model (themselves spatially autocorrelated) can account satisfactorily for the observed spatial autocorrelation in the dependent variable. In such a situation, regression residuals generally are found to be negligibly autocorrelated, and standard regression approaches are adequate. At other times, the researcher might introduce a variable or variables that capture the influence of the geophysical or other forces underlying the spatial effect. Fotheringham, Brunsdon and Charlton provide several examples – GWR among them – of how this particular issue might be approached (2002, p. 15-24). As a general matter, it is wise practice to model, perhaps with a simple regression specification, the heterogeneity of a spatial process before spatial dependence modeling is undertaken. The reason for this is that spatial dependence modeling assumes a homogeneous (technically, stationary) spatial process.

If, on the other hand, the association is an interaction suggesting some type of formal dependency among areal units, then a modeling strategy with a spatially dependent covariance structure is the way to proceed. In this instance, controlling for heterogeneity likely will not fully remove the spatial effects within the data. An alternative is needed – a spatially oriented approach that formally incorporates a spatially lagged dependent variable or spatially lagged error term. In a conceptual way, this approach is a spatial analogue to the treatment of temporal variables in time series analysis. The added dimensionality of geographic space and the absence of any form of natural order in spatial data, however, render many statistical procedures in time series analysis inappropriate in spatial analysis. For details on spatial dependence modeling, the reader is advised to begin with Anselin (1988), Anselin & Bera (1998), and Anselin (2003). This literature presently is expanding at a very rapid pace.

Concluding our discussion of population change in the Great Plains, we attempt to bring several of these thoughts together by presenting some regression results in Table 1. The table has four columns of regression parameter values and some useful diagnostic terms. In the table, we take the dependent variable (log population growth in the 1990s) and regress this on a few independent variables. The first column shows the results of a standard OLS regression. We take initial satisfaction in noting that the OLS model performs reasonably well. Several parameter estimates are strongly significant, parameter signs are as anticipated, and the adjusted R^2 statistic achieves a respectable level of 0.337. Having anticipated and checked for it, however, we quickly note a problem: the regression residuals show strong spatial autocorrelation (Moran's $I = 0.363$), a clear indication that the model is in violation of at least one of the assumptions underlying standard linear regression. The Moran test tells us that the residuals are not independent. Moreover, the Koenker-Bassett test for heteroskedasticity indicates that the residuals also are not distributed identically. Both are serious violations of OLS assumptions and suggest that inferences drawn from the model in column one could be seriously flawed.

[Table 1 About Here]

Comparing the residual spatial autocorrelation ($I = 0.363$) with the spatial autocorrelation for the dependent variable (reported above, $I = 0.542$) tells us that spatial autocorrelation in one or more of our independent variables actually "explains" a portion of the spatial autocorrelation in the dependent variable. As indicated above, it frequently is the case that the independent variables in a regression model can almost completely account for the spatial autocorrelation in a dependent variable, thus removing a problematic spatially autocorrelated residual. However, in our case, the regressors have not satisfactorily accounted for obvious spatial heterogeneity and/or dependence in the data, and a correction to the model clearly is indicated. But what type of correction? At this point one's theory of the process under investigation is asked to do some heavy lifting. Does the residual dependence in the model likely stem from omitted variables on the right-hand side of the regression specification, thus suggesting the utility of a spatial error model? If so, we might pause to ask, what variables? On the other hand, might there be spillover influences among growing counties or declining counties that directly influence the growth rates of their neighbors? Fortunately, to supplement our theory about the process (however strong), we receive some additional guidance from other diagnostic statistics applied to the residuals in the OLS regression.

Two such regression diagnostics are shown at the bottom of the first column: Lagrange Multiplier test statistics which, for this example, suggest a preference for a spatial lag specification (a lagged dependent variable term on the right-hand side) over a spatial error specification (a lagged error term). While often very helpful, these diagnostic statistics are also known to be unreliable in the presence of unresolved heterogeneity in the model. We therefore have added a second-order polynomial trend surface to the OLS model in the hope of capturing at least a portion of the spatial heterogeneity in the data.

Using ESDA software we examined the shape of the north-south and east-west marginal distributions of the dependent variable, and on that basis we chose a second-order trend surface and added to our OLS model five variables expressing linear and nonlinear aspects of the geographic centroid of each county: latitude, longitude, latitude-squared, longitude-squared, and latitude-x-longitude (results in column two of Table 1). Few of the parameters of the added variables are statistically significant. Understandably, there is a correlation between latitude (a north-south variable) and the temperature range variable. The addition of the geographic variables thus reduces the significance of the latter, and latitude and its square are not significant contributors with the temperature variable in the model. Other parameters also change in the shift from column one to column two. Quite interestingly, after controlling for the geographic variables, county acreage assumes significance in the anticipated direction.

Yet the model in column two remains unsatisfactory. Moran's *I* is modestly reduced, but it and the heteroskedasticity test both suggest the need to deal with spatial dependence. The model, when augmented with the trend surface variables (column two), unambiguously suggests dependence in the form of a spatially lagged dependent variable. Yet heteroskedasticity remains high, thus reducing our confidence in the Lagrange Multiplier tests. Consequently we present both a spatial error model (column three, Table 1) and a spatial lag model (column four), partly as a concession to our uncertainty about the process but partly also to see what additional understanding we might glean from examination of both the spatial error and lag specifications.

We comment first on the results of the spatial error model shown in column three. In this specification, the error variance-covariance matrix is assumed to have non-zero off-diagonal terms (thus permitting the extent of autocorrelation in the errors to be estimated by a parameter, λ). The underlying assumption in the model (apart from those assumptions justifying a linear specification and the particular set of selected independent variables) is that spatial autocorrelation in the dependent variable is caused by one or more spatially autocorrelated "omitted variables" on the right-hand side of the regression specification. Such a specification is often appropriate in the absence of a theoretical rationale for assuming interaction dependence in the dependent variable. If indeed a spatial error specification is the "correct" specification for the process, then the estimated parameters from the OLS regression are unbiased but inefficient (the standard errors of the parameter estimates are downwardly biased in the presence of positive spatial autocorrelation). We note that the parameter estimates in column three (compared to column two) are modestly different. Most notably, both the initial population and acreage variables lose their significance, and the only remaining strong substantive parameter is that for our key independent variable (farm employment). A higher likelihood and lower (i.e., more negative) AIC (Akaike Information Criterion) score in column three are encouraging, but the pseudo R^2 statistic is considerably lower than achieved in the OLS runs. We note two additional desirable features of this model: the spatial error parameter (λ) is strong, and the model has eliminated any diagnostic evidence of a remaining spatial lag influence (i.e., the Lagrange Multiplier test for remaining lag specification is small and not statistically significant. Aside from the remaining heteroskecasticity, the model appears to be a plausible alternative to the OLS specification.

We now comment on the spatial lag model shown in column four, a model we anticipate from the OLS diagnostics to be the appropriate model. In this specification, a lagged version of the dependent variable appears on the right-hand side of the regression specification. As discussed above, the particular form of the spatial lag is determined by the researcher through a definition of "neighborhood," operationalized by a weights matrix. The strength of the lag effect is estimated by the lag parameter, ρ. A spatial lag specification is particularly appropriate when there is structured spatial interaction involving the dependent variable and when the analyst is concerned about measuring the strength of that interaction or is concerned about having "correct" estimates of the regression parameters which can be obtained only after removal of the effect of spatial autocorrelation in the process. If a spatial lag model is the "correct" specification for the data-generating process at hand, then the incorrect OLS specification will suffer from biased and inconsistent parameter estimates.

We note modest changes in the estimated regression parameters, including the observation that inclusion of the lag term has removed the importance of the initial population variable but has not reversed the sign of this parameter (as occurred with the spatial error model). The spatial lag parameter (ρ) is strong and significant and, of the four models, this specification has the highest likelihood and lowest AIC score. Some indication of residual error correlation is apparent in this model, and that is of some worry. It suggests that a model including both a lag and error specification may yet be a preferred fifth model. But we do not pursue that route here. Our inclination at this point is to state a preference for the lag specification over the error specification, and we are not at all uncomfortable with the implied theoretical position that sprawl and residential spillover growth into neighboring counties (and, elsewhere, spillover influences of population loss) are likely the source of difficulty in the OLS misspecification.

SUMMARY

In this chapter, we have discussed the role of geographic space in quantitative demography. A re-emerging interest in spatial demography is beginning to appear as an increasing number of demographers seek to adopt the formal tools of spatial econometrics to improve on traditional regression models of demographic processes operating in space. The concept of spatial autocorrelation and ways to specify correctly multiple regression models in the presence of spatial autocorrelation are made more concrete through an illustration of spatial modeling of county-level growth in the U.S. Great Plains region during the 1990s.

It is our belief that as our own statistical models become more sophisticated, as spatial processes are brought into empirical demographic studies to correct for potential misspecification, and as our work begins to add in significant ways to the larger literature on spatial data analysis, we will have moved the science of spatial demography forward in very exciting ways. The growing interest in the field of spatial econometrics among several disciplines in the social sciences, of which the re-emergence of interest in spatial demography is a part, suggests an exciting future for quantitative demographers.

Paul R. Voss
Department of Rural Sociology
University of Wisconsin-Madison

Katherine J. Curtis White
Center for Demography and Ecology
University of Wisconsin-Madison

Roger B. Hammer
Department of Rural Sociology
University of Wisconsin-Madison

NOTES

Please direct all correspondence to Paul R. Voss at 316 Agriculture Hall, 1450 Linden Drive, Madison, WI, 53706, or voss@ssc.wisc.edu. The authors extend their appreciation to David Long and Nick Fisher for assistance and advice regarding the GIS applications and spatial modeling for the Great Plains working illustration, to Jeremy White for graphic support, and to the editors and Glenn Deane for their extensive comments on earlier drafts. This research was supported in part by the U.S. Department of Agriculture, Hatch Grant WIS04536, by the National Institute for Child Health and Human Development, Center Grant HD05876 and Training Grant HD07014, by the University of Wisconsin Center for Demography and Ecology, through its Geographic Information and Analysis Core.

REFERENCES

Anselin, L. (1988). *Spatial econometrics, methods, and models*. Dordrecht: Kluwer Academic.

_____. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, 27,93-115.

_____. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M. Fischer, H. J. Scholten, and D. Unwin (Eds.), *Spatial analytical perspectives on GIS*, (pp. 111-125). London: Taylor & Frances.

_____. (2001). Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference*, 97,113-139.

_____. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review*, 26(2),153-166.

Anselin, L. and Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah and D. Giles (Eds.), *Handbook of applied economic statistics*, (pp. 237-289). New York: Marcel Dekker.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*. Harlow Essex: Longman Scientific & Technical.

Brueckner, J. K. (2003). Strategic interaction among governments: an overview of empirical studies. *International Regional Science Review*, 26(2),175-188.

Chou, Y.- H. (1997). *Exploring spatial analysis in Geographic Information Systems*. Santa Fe, NM: OnWord Press.

Cliff, A. D. and Ord, J. K. (1973). *Spatial autocorrelation*. London: Pion Limited.

_____. (1981). *Spatial processes: models and applications*. London: Pion Limited.

Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: Wiley.

Deane, G. E., Beck, E. M., and Tolnay, S. E. (1998). Incorporating space into social histories: how spatial processes operate and how we observe them. *International Review of Social History*, Supplement 6 43:57-80.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. E. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: Wiley.

Griffith, D. A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In S. L. Arlinghaus (Ed.), *Practical Handbook of Spatial Statistics*, (pp. 65-82). Boca Raton: CRC Press.

Haining, R. (2003). *Spatial data analysis in the social and environmental sciences*. Cambridge: Cambridge University Press.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis.* New York: John Wiley & Sons.

_____. (1985). *Exploring data tables, trends, and shapes*. New York: John Wiley & Sons.

Lobao, L. (2004). Continuity and change in place stratification: Spatial inequality and middle-range territorial units. *Rural Sociology*, 69(1),1-30.

Lobao, L. and Saenz, R. (2002). Spatial inequality and diversity as an emerging research area. *Rural Sociology*, 67(4),497-511.

Martin, R. J. (1987). Some comments on correction techniques for boundary effects and missing value techniques. *Geographical Analysis*, 19,273-282.

Messner, S. F. and Anselin, L. (2004). Spatial analyses of homicide with areal data. In M. F. Goodchild and D. G. Janelle (Eds.), *Spatially Integrated Social Science*, (pp. 127-144). Oxford: Oxford University Press.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37,17-23.

Rogers, E. M. (1962). *Diffusion of innovation*. New York: The Free Press.

Stephan, F. F. (1934). Sampling errors and interpretations of social data ordered in time and space. *Journal of the American Statistical Association*, 29(185 Supplement),165-166.

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

Tolnay, S. E., Deane, G., and Beck, E. M. (1996). Vicarious violence: spatial effects on southern lynchings, 1890-1919. *American Journal of Sociology*, 102(3),788-815.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing.

Voss, P. R. (1993). Applied demography and rural sociology. In D. L. Brown, D. R. Field, and J. J. Zuiches (Eds.) *The demography of rural life*, (pp. 145-170). University Park, PA: Northeast Regional Center for Rural Development.

Voss, P. R. (2004). Demography as a spatial social science. Paper presented at the Annual Meeting of the Southern Demographic Association.

White, K. J. C. (2003). A century of population change in the U.S. Great Plains. Ph.D. dissertation, University of Washington.

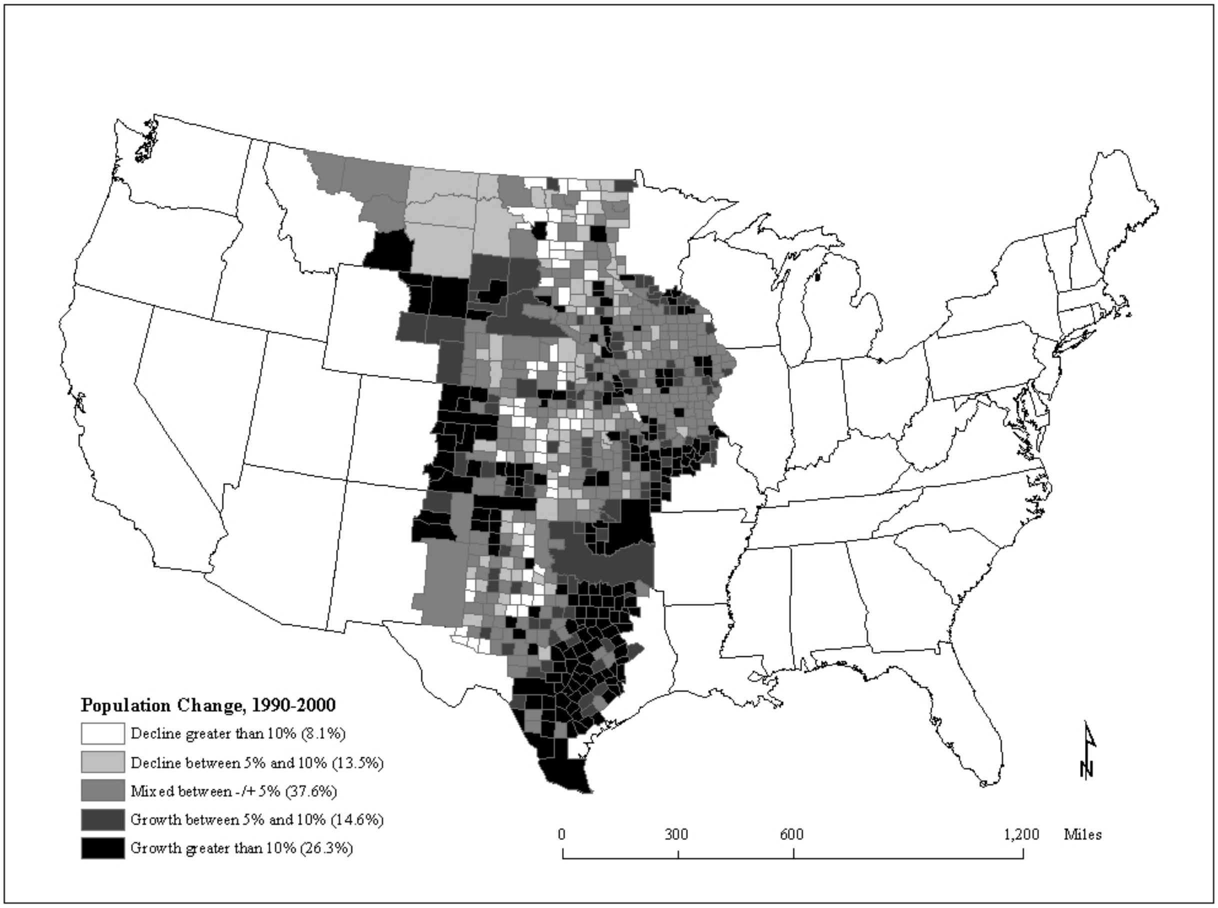
Population Change, 1990-2000
Decline greater than 10% (8.1%)
Decline between 5% and 10% (13.5%)
Mixed between -/+ 5% (37.6%)
Growth between 5% and 10% (14.6%)
Growth greater than 10% (26.3%)
0
300
600
1,200
Miles
N

Figure 1. Spatial Distribution of Population Change among Great Plains Counties, 1990-2000

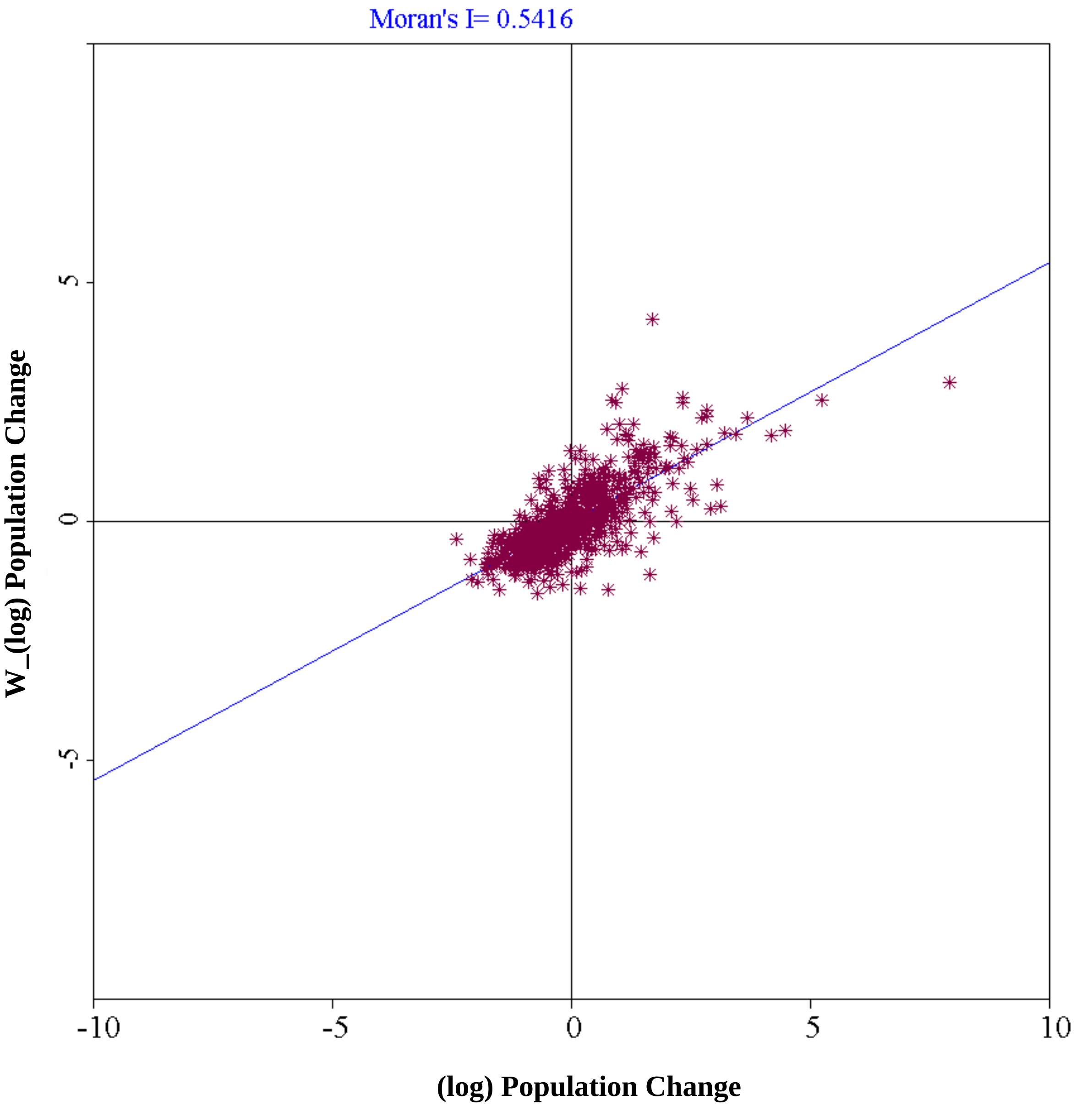

Figure 2. Moran Scatterplot of Population Change

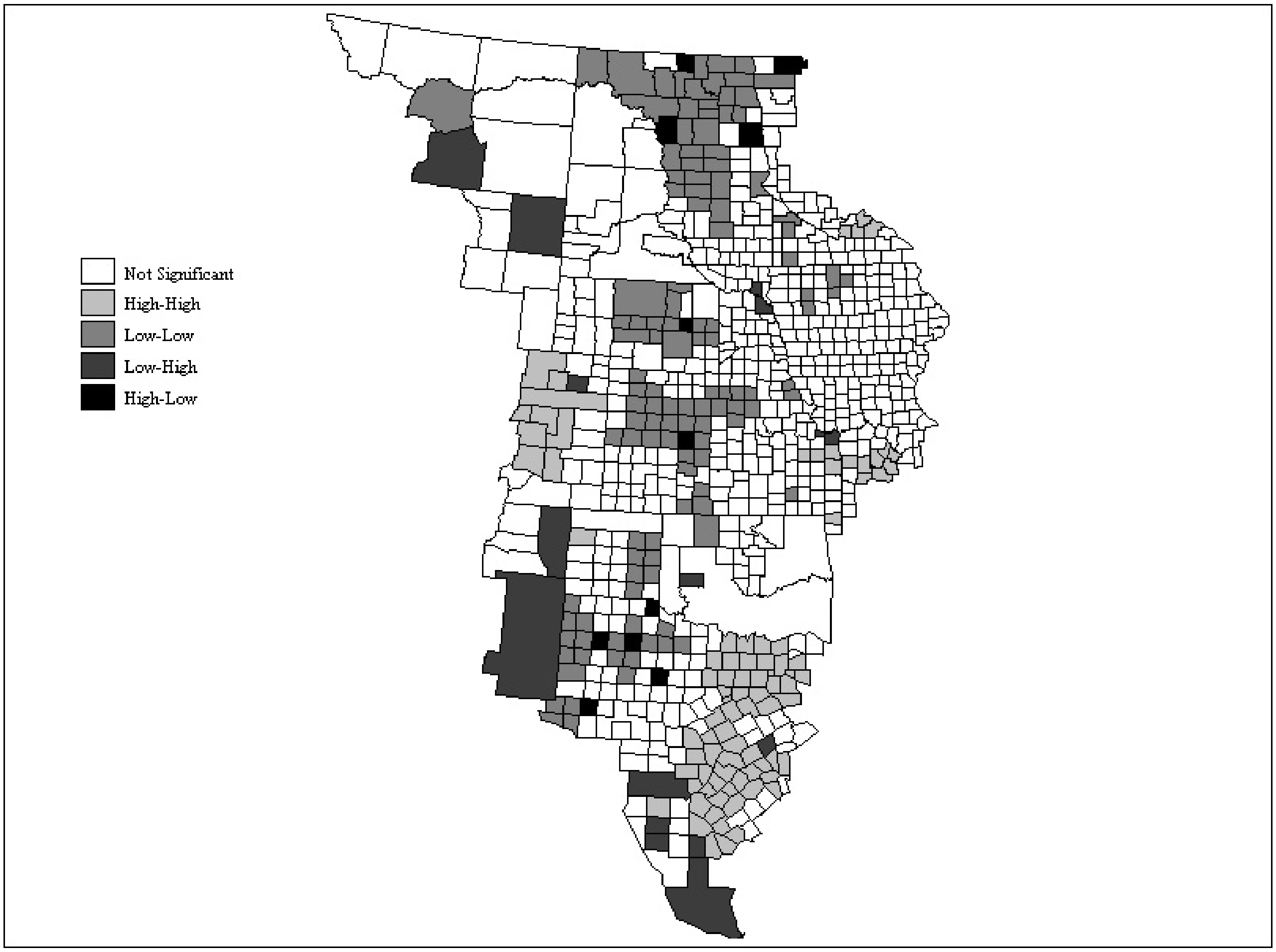

Figure 3. LISA Cluster Map of Population Change

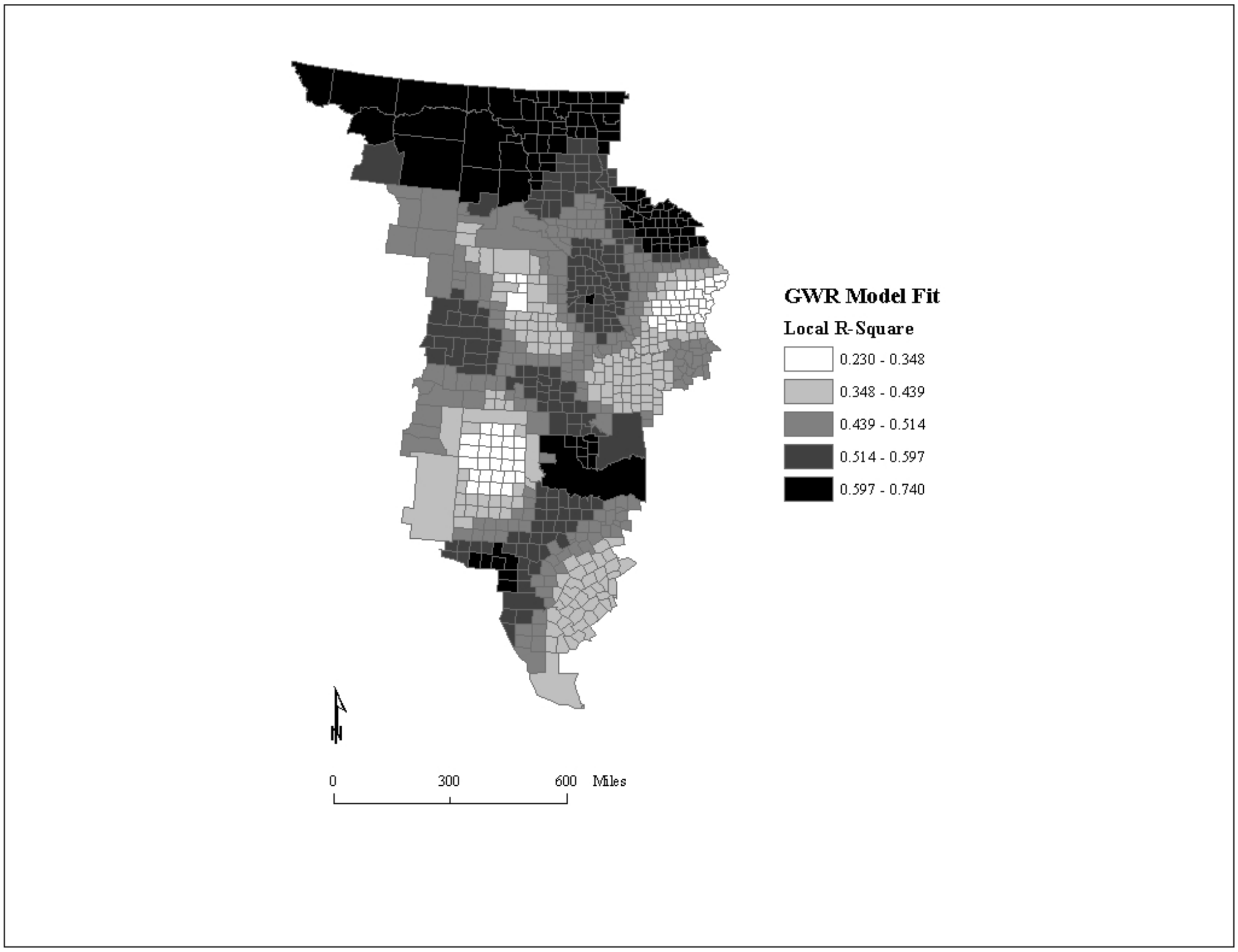

Figure 4. GWR Derived Distribution of Local R^2 Estimates

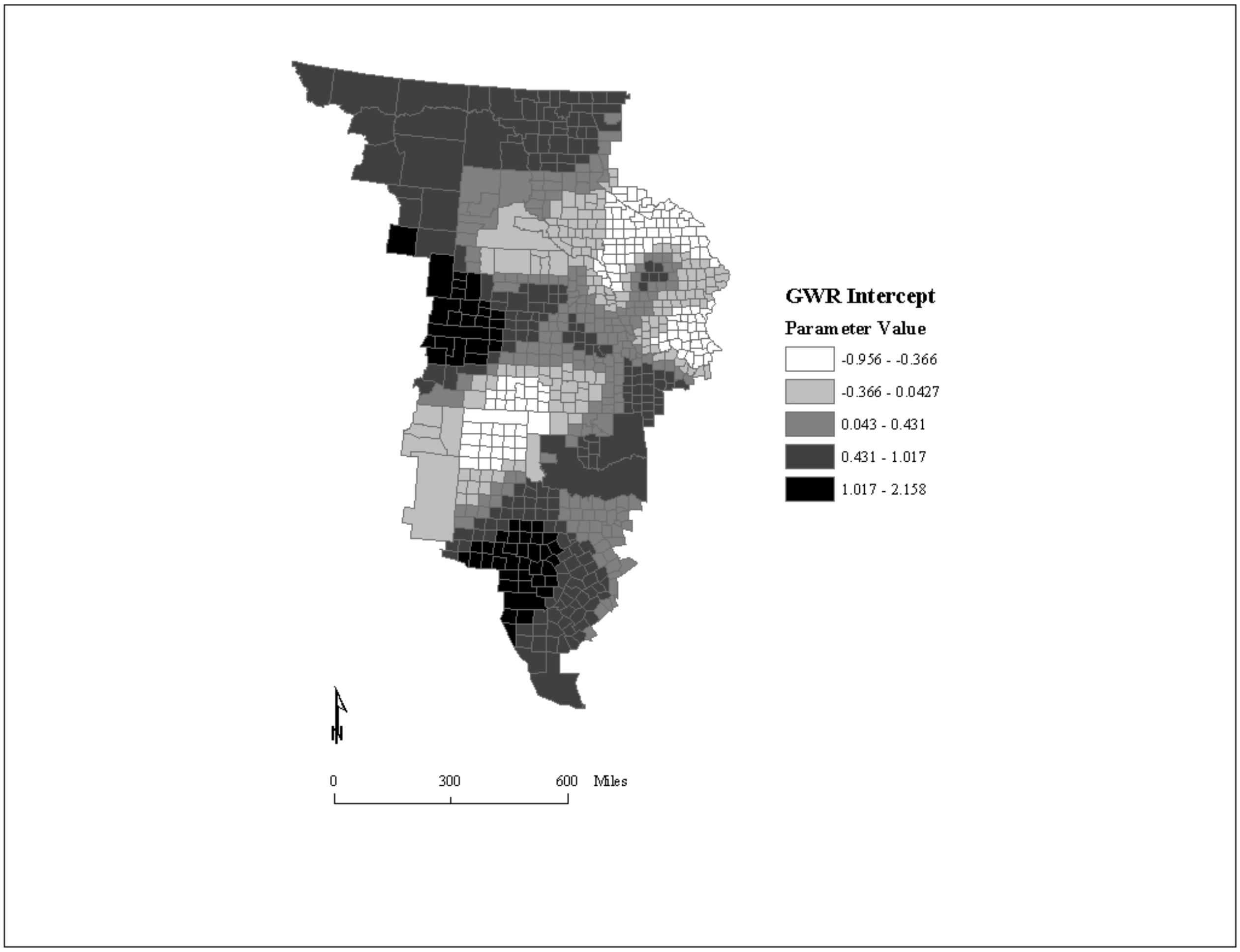

Figure 5. GWR Derived Distribution of Intercept Parameter Estimates

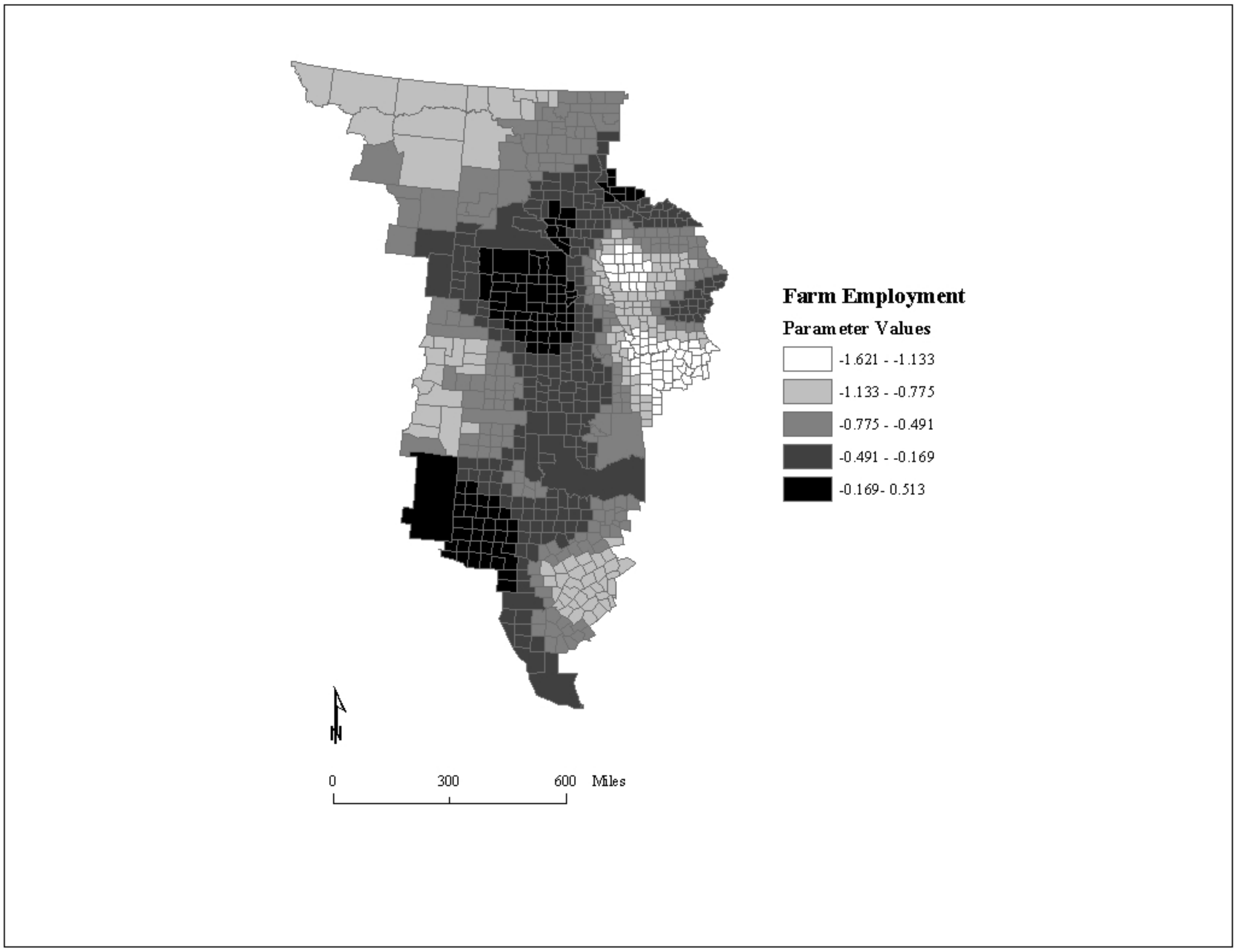

Figure 6. GWR Derived Distribution of Farm Employment Parameter Estimates

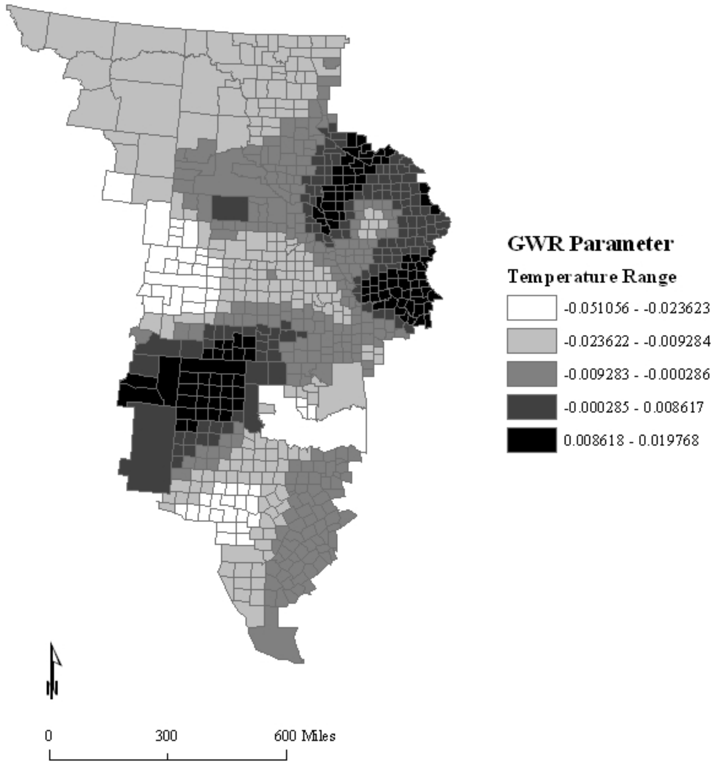

Figure 7. GWR Derived Distribution of Temperature Range Parameter Estimates

Table 1. Non-Spatial and Spatial Regression Models of County Population Change, 1990-2000

Dependent Variable = log (Percent Population Change)

	(1)	(2)	(3)	(4)
		Geographic		
	OLS	Coordinates	Spatial Error	Spatial Lag
Farm Employment (β)	-0.332 ***	-0.342 ***	-0.459 ***	-0.420 ***
	(0.092)	(0.096)	(0.081)	(0.078)
Proportion of Pop < 18 (β)	0.190	0.223	0.096	0.091
	(0.123)	(0.120)	(0.111)	(0.097)
Temperature Range (β)	-0.008 ***	-0.006 *	-0.005	-0.004
	(0.001)	(0.002)	(0.003)	(0.002)
City Status (β)	0.014	0.007	0.011	0.012
	(0.012)	(0.012)	(0.010)	(0.001)
log County Acreage (per 100,000) (β)	-0.004	-0.026 **	-0.006	-0.008
	(0.007)	(0.010)	(0.009)	(0.008)
log Initial Population (per 1,000) (β)	0.025 ***	0.026 ***	-0.001	0.004
	(0.006)	(0.007)	(0.006)	(0.006)
Intercept (β)	0.309 ***	0.041	0.011	0.020
	(0.049)	(0.076)	(0.081)	(0.061)
Latitude		0.00002	0.00003	0.00002
		(0.00001)	(0.00002)	(0.00001)
Latitude^2		0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)
Longitude		0.0002 *	0.0002	0.0001 *
		(0.00007)	(0.0001)	(0.0005)
Longitude^2		0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)
Latitude*Longitude		0.000 ***	0.000 ***	0.000 ***
		(0.000)	(0.000)	(0.000)
Spatial Error Paramter (λ)			0.679 ***	
			(0.035)	
Spatial Lag Parameter (ρ)				0.651 ***
				(0.034)
Adjusted or Pseudo R^2	0.337	0.378	0.280	0.502
Moran's I (error)	0.363 ***	0.341 ***		
Likelihood	619	646	751	765
AIC	-1225	-1268	-1478	-1503
Heteroskedasticity	23.505 *** †	51.067 *** †	328.807 *** ‡	335.155 *** ‡
Robust Lagrange Multiplier (error)	2.895	0.327		10.924 **
Robust Lagrange Multiplier (lag)	322.013 ***	41.174 ***	0.106	

* p < .05, ** p < .01, *** p < .001 (two-tailed tests)
† Koenker-Bassett Test for Heteroskedasticity, ‡ Breusch-Pagan Test for Heteroskedasticity