Movement in America: A Spatial Data Analysis of County-to-County Migration

Christopher Inkpen
The Pennsylvania State University

Abstract

Internal migration is a dynamic socio-geographic process that impacts individuals, families, and communities. Given the lower inertia and uncertainty involved in moving, internal migration in the United States can be seen as a measure of how actors in fluid labor markets respond to economic shocks. Yet social ties and demographics may also influence migration. This study uses county-level data from the U.S. Census Bureau's American Community Survey 5-year estimates from 2008 to 2012 to answer two interrelated questions; (1) what are the constituent determinants of county-to-county migration in the contiguous United States and (2) what is the spatial structure of this county-to-county migration? Employing spatial regression methods, I examine the relationship of county-to-county migration and a number of demographic and economic indicators. I find that county-to-county internal migration is a highly spatially structured process, and that accounting for spatial structure reduces the impact of important migration predictors while increasing others.

Introduction:

Internal migration is a dynamic geographic and social process in a country that can affect individuals, families, and communities. Macro-level economic forces push and pull individuals and families across state and county borders. Lee classifies migration as an economic decision, based on positive and negative features of the area of origin and destination combined with intervening obstacles and personal factors (Lee, 1966). De Jong characterizes the decision to migrate as determined by individual human capital attributes, household characteristics, and community characteristics (De Jong, 1999). Other scholars assert the social component of migration, noting that while migration is used as a household strategy to improve economic mobility, the act of migration may create feedback mechanisms that lead to a self-perpetuating migration stream as migrant networks become institutionalized (Massey et al., 1990).

Given the lower respective risk and obstacles of migrating internally as opposed to internationally, coupled with the greater availability of labor market information on domestic destination spots, internal migration in the United States should be seen as a measure of how actors in fluid labor markets respond to shocks. Yet social ties and perceived areas of acceptance may drive migration in the absence of clear economic opportunities. When focusing on county-to-county migration, previous studies find poverty associated with in-migration, due to rural-to-rural networks of poor movers seeking employment (Nord, 1998; Ravallion and Wodon, 1999). Moreover, factors that influence migration both within and outside of a state may be distinct from those that impact solely state-to-state migration. Migration for single adults may also be different than migration for families. Thus, migration can be seen as a complex phenomenon with myriad factors that influence individuals and families to prioritize certain destination areas over others.

As an inherently geographic process, internal migration is a potential candidate for spatial analysis, as places that people move to are likely to be clustered on certain characteristics. This study uses county-level aggregate data from the United States Census Bureau's 5-year estimates American Community Survey from 2008 to 2012 to answer two interrelated question; (1) what are the constituent determinants of total county-to-county migration in the contiguous United States and (2) what is the spatial structure of this county-to-county migration? Employing spatial regression methods, I examine the relationship between internal migration between counties and the ratio of children and elderly to the working population (dependency ratio) along with a number of economic and demographic variables. I find that county-to-county internal migration is a highly spatially structured process, and that including spatial coefficients reduces the impact of important migration predictors while increasing others.

Research Design:

In order to explore the spatial structure of internal migration, this investigation uses the tools of exploratory spatial data analysis to underscore where internal migration is concentrated at both high and low levels. I explore the spatial clustering of internal migration first by examining maps of the internal migration rate as well as the dependency ratio. This study then tests various specifications of a spatially configured weights matrix to explain how neighboring counties and clustering influence migration. This weights matrix is used to produce exploratory spatial data measures of Moran's *I* and Local Area Moran's *I* statistics to gauge whether the univariate spatial distribution of internal migration and the dependency ratio is statistically significant. This analysis employs ordinary least squares regression as a preliminary test and spatial regression models to answer two key questions: (1) How does the dependency ratio of counties relate to county-to-county migration in the United States? (2) What is the spatial structure of county-to-county internal migration in the United States? I propose three hypotheses as potential answers these questions:

1. The dependency ratio is negatively related to internal county-to-county migration, and will be significant despite the introduction of spatial coefficients.
2. Economic factors within counties will attenuate the relationship between the dependency ratio and internal migration.
3. The error structure of a standard linear model regression will be spatially correlated, requiring spatial regression models to account for spatial clustering on the dependent and independent variables.

Data and Measures:

This inquiry uses county-level aggregate data from the American Community Survey (ACS) 5-year estimate from 2008 to 2012 to examine internal migration and its constituent determinants in a post-recession America. The ACS compiles a number of questions regarding the socio-economic and demographic characteristics of the residents of the United States. Although individual-level samples are available, the questions that are released at the individual, county, and state levels vary. In order to measure county-to-county migration, I utilize residence questions to determine internal migrant stocks within counties. The ACS 5-year estimate (as opposed to other census products) asks its respondents whether they are living in the same house as the previous year. When the answer is no, they are prodded as to whether or not they are living in the same county as the previous year. For the 2008-2012 5-year estimate, 5% of the pool (or 16.6 million respondents) reported that they changed both their house and their county of residence. Moreover, these respondents report if they moved from within the state to their new county or if they arrived from a different state. These measures are decomposed but are combined in this study to create a total internal county-to-county in-migration rate (excluding new foreign born migration). However, due to data confidentiality issues, the census only releases this information at the aggregate level for counties. Thus, this data set is composed of the 3,109 counties in the contiguous United States (excluding Alaska, Hawaii, Puerto Rico, and Guam to allow for the use of spatial contiguity matrices).

This study examines county-level measures for;

Dependent Variable

a) total county-to-county migration rate

Independent Variables

b) the ratio of children and elderly to the working age population (dependency ratio)

c) the percent of county-residents living below the poverty level

d) the percent of unemployed county-residents aged 16 and up

e) the population of the county (divided by 1000)

f) the average percentage of annual household wage increase from 2008 to 2012

g) the median monthly housing costs

h) 9 U.S. census regional division variables (described below)

i) the number of counties in a state

j) the USDA typology code for "mining" counties

k) the USDA typology code for "retirement" counties

l) 5 county-type profile variables based on dependency ratio, population, and average annual per capita earnings increases.

In this analysis, the county-to-county internal migration rate is calculated by dividing the number of respondents who reported a county-to-county move (either within or from outside of the state) by the population of the county of residence and multiplying by 1000. The dependency ratio is calculated by summing the number of county residents ages 0 to 14 as well as summing the number residents aged 65 and older. These figures, added together, are then divided by the county population aged 15 to 64. This creates a proportion of the young and old to the working population. In this analysis, the dependency ratio variable is transformed by multiplying the proportion by 100 for easier interpretation of regression model coefficients. Thus, greater values on the dependency ratio indicate a greater proportion of the population that is both very young and retirement age.

Distinct from labor-market opportunities and unemployment, this study also utilizes the percent of respondents living below the poverty line to gauge overall county poverty. I include the unemployment rate per county by dividing the number of county-residents over 16 currently looking for work but unemployed by the total number of working-age residents. County population in thousands is also included.[1] To account for the one aspect of the county-specific economic impact of the recession, this study takes average household wages in 2008 and 2012 and measures the percentage of increase or decrease during that time. In order to capture whether larger states attract more internal immigrants, I include the number of counties in each state as a control variable. Moreover, to measure housing prices, the models specify the median housing costs within a county. Finally, this study incorporates two of the United States Department of Agriculture (USDA) county typology dichotomous variables to determine counties with economies based on mining as well as counties that are traditionally retirement locations. These typological distinctions may explain traditional or contemporary causes of internal migration in the United States (retirement and expansion in the mining industry).

In order to capture regional differences that attract movers, this analysis includes the U.S. Census Bureau's divisional classifications as dummy variables. The distribution, as shown in detail in the appendix, describes the regional distribution of counties as follows; New England, Mid-Atlantic, East North Central, West North Central, South

[1] Population density as well as a dichotomous variable for metropolitan area (derived from the United States Department of Agriculture rural-to-urban continuum codes) were included in initial models in lieu of population but omitted due to limited explanatory power.

Atlantic, East South Central, West South Central, Mountain, and Pacific.[2] These independent variables serve to capture a portion of the socioeconomic and demographic variability within the United States related to internal migration.

In addition to socio-demographic variables, this investigation uses a clustering algorithm to group counties into similar clusters based on their dependency ratio, population, and the U.S. bureau of commerce's percent change of per capita personal income from 2007 to 2008 (during the recession) and the percent change of per capita personal income from 2011 to 2012 (after recovery). This grouping of variables seeks to describe how counties with different demographic profiles (both in dependency ratio and population size) responded to the Great Recession. These group indicators are then included in the analytical models.

Grouping analysis in this study is achieved using the K-means clustering algorithm ArcGIS 10 to find natural structure within the dataset based on the specified variables. As an important distinction, the K-means algorithm does not impose spatial constrains and instead seeks to group counties based on their Euclidean distance (on a hyper-plane of the included features) from a pre-set number of centroids. This algorithm uses randomly selected seeds (spatially weighted to favor selection of distant seeds) to create the clusters. Observations are sorted into the closest grouping and group-mean data centers are computed iteratively until group membership converges at a stable number. Furthermore, the clustering analysis tool in ArcGIS 10 allows for group cluster optimization, which uses the Calinski-Harabasz pseudo F-statistic (a ratio of within-group similarity to between-group differences) to calculate the optimal number of clusters based on the selected features. The results of this clustering algorithm (found in appendix in Figure 1) suggest by a wide margin the use of five distinct groups that differ

[2] Distribution of states in divisional variables are listed in appendix.

on their composition of dependency ratio, population, and average per capita income increases both during and after the recession. These groups are described in detail below.

Descriptive Statistics

Table 1 reflects the distributions of the dependent and independent variables used in this study. The summary measure of central tendency and dispersion show that the average county received approximately 62 new people per thousand residents as a result of internal migration in the United States. Noting the minimum and maximum values of 0 and 374, it is clear that counties experience internal migration differently across the United States. In fact, the 75 percentile figure of 74 internal county-to-county migrants per thousand residents indicates, and the univariate histogram in figure 1 confirms, that the variable is highly right-skewed, with the majority of counties experiencing county-to-county internal migration as under 70 new people per thousand whereas some outliers receive nearly 1/5th of their county residents as county-to-county migrants. Figure 2 is a quintile map of the spatial distribution of county-to-county migration rate across the United States. Visually, it is evident from this spatial map of the dependent variable that county-to-county migration is clustered highest in Texas, the Midwest, counties bordering California, as well as in parts of northern Florida and North Carolina.

Table 1 and Figure 1 about here

The dependency ratio, while transformed, shows the proportion of dependent residents to the number of working age residents. A mean of 69 and a 75 percentile value of 74 confirm that the majority of counties have greater numbers of working-age residents than children and elderly. This variable, seen as a histogram in figure 3 (located in Appendix), is fairly normally distributed but has long tails that account for mainly working populations and populations with a great deal of older residents. The quintile map of dependency ratio shown in figure 4 (located in the appendix) reveals clustering again in Texas but shows greater concentrations of dependency through the Great Plains and the Southwest. Southern Florida, as expected, also shows high levels of dependency ratio given their elderly internal migrant population. Percent of poverty per county lists a mean of 15% and a 75th percentile figure of 17%, indicating a right-skewed distribution, although there are outliers listing 0 percent living in poverty (richer counties) and those that list up to 46% living in poverty (highly impoverished counties). The unemployment rate also follows this right-skewed pattern, with a mean of 8.6% unemployed working age residents per county and 50% of the counties falling between 6% and 10.5%. Household wages in counties increased by 5%, and although there is great variability in this measure with, 50% of the counties fall between a 1 to 9% increase, indicating that most counties have experienced growth in wages between 2008 and 2012. Median monthly housing costs range from $540 at the 25th percentile to $830 at the 75th percentile, with an average of roughly $720 per month. This may indicate reasonable variability in housing for most counties. Noting the USDA typology county codes, only 5% of counties are listed as mining counties whereas 14% are traditional retirement spots.

The five county profile groups created using the K-means clustering algorithm vary on a number of socio-demographic and economic indicators (shown in Table 2 and Figure 5). Cluster 1 (comprised of 1,336 counties, 43% of the data) closely approximates large towns, with an average population of approximately 35,000, 9% unemployment, and 2/5th of the population with college degrees. Notably, cluster 1 has the highest percentage of county population living in poverty (at 16%) and the lowest proportion with college degrees. Cluster 2 (1,354 counties, 44% of the data) mirrors closely the demographic profile of small cities, with an average population of 138,000, 15% of population living in poverty, and 4% household wage increases between 2008 and 2012. Cluster 3, as noted by its small group size (28 counties) but large average population (2.5

million) matches the demographic profiles of large cities. This is evidenced in Figure 5, as the counties surrounding Los Angeles, New York, Philadelphia, Chicago, and Miami all pertain to Cluster 3. This cluster has the lowest dependency ratio (60), second highest percentage of college graduates (47%), and the highest percentage of foreign born population (25%) but also the lowest internal migration rate (at 40 per 1000 people) and the lowest increase in household wages (2%). Cluster 4, as seen in Figure 5, parallels small towns in the Great Plains, with the highest dependency ratio (80), lowest average population (11,000). Interestingly, counties in cluster 4 experienced an overall loss in per capita personal income from 2007 to 2008 of 3.2% but experienced an increase between 2011 and 2012 of 12.5%. Moreover, with the 2nd highest household wage change from 2008 to 2012 of 9.3%, counties in the Great Plains small town cluster rebounded from the recession. Lastly, cluster 5 is unique in that it has the highest percentage of college graduates on average (47.1%), the lowest percentage of foreign-born population (2.7%), and the greatest change in annual household wages from 2008 to 2012 (11.6%). Moreover, change in per capita income during the recession was notably positive (18.2%) and remained highly positive after recovery (23.3%). Given the location in the West North Central division (including the Dakotas and Minnesota), this cluster most likely represents counties that experienced expansion in the energy sector that spurred economic activity.

Table 2 and Figure 5 about here

A piece-wise correlation matrix of the dependent and independent variables (omitted due to size) demonstrates that the dependent and independent variables do not exhibit a high degree of colinearity, with the highest correlation coefficient shared between continuous variables being percent unemployed and percent in poverty (0.59). Clusters 1 and 2 are inversely correlated at fairly high level (-0.76) and cluster 3, as a profile of cities, is correlated with population at 0.71. Figure 6 is a bivariate scatter plot describing the negative relationship (-0.23 correlation coefficient) between county-to-county migration and the dependency ratio, which mirrors the first hypothesis. As a whole, these figures briefly summarize the distributions of the variables of interest, yet simple quintile maps reveal evident regional clustering in our dependent and independent variable. Thus, a deeper look using spatial statistics is required to understand how these variables are nationally distributed.

Figure 6 about here

Methods and Exploratory Spatial Results

Spatial Weights Matrices and Descriptive Spatial Statistics

Spatial statistical analysis requires the creation of a spatial weights matrix that defines how and how many neighboring units can influence a central actor in order to account for spatial autocorrelation on variables (Anselin, 1988). This study tests three weights matrices; (1) a Rook's 1st order matrix, (2) a Queen's 1st order matrix, and (3) a K-nearest neighbors matrix using 4 neighbors. Table 1 shows the Moran's *I* statistics for the dependent and independent variables using all matrices. A Queen's 1st order matrix describes a neighbor as any unit that shares a common point (border or boundaries), whereas a Rook's 1st order contiguity matrix requires that the units share at least some positive portion of a boundaries is shared (not simply a point) (Kelejian and Prucha, 2010). In this study, Figure 7 represents a histogram of the number of neighbors for each county, indicating a modal response of 6 neighbors and a range of between 2 and 14 county-neighbors. The *k*-Nearest neighbor distance matrix uses the four closest neighboring units based on county centroids to create neighbors and often provides a solution to distance-based weights in the presence of unit area variety (Chi and Zhu, 2007). I create these three weights matrices to generate the spatial statistics of Moran's *I* and Local Indicator of Spatial Association (LISA) values. Along with limitations in certain spatial regression models, I use these values to compare differences in spatial autocorrelation among the variables of interest to inform the selection of the weights matrix.

Moran's *I* is an indicator of the statistical association between a value of interest for a unit and the average value of the unit's neighbors as defined by the weights matrix (Moran, 1950). Positive numbers on Moran's *I* indicates spatial clustering of similar attribute values. Table 1 shows that in this model, regardless of the weights matrix used all dependent and independent variables are significant at below the 0.001 level, indicating the presence of spatial correlation for all the variables of interest.[3] The dependent variable of county-to-county migration has a value of roughly 0.18 in all models, whereas dependency ratio is even more spatially autocorrelated with a general value of between 0.41 and 0.47. Percent of unemployed county population above the age of 16 appears to be a fairly spatially autocorrelated variable, with values ranging from 0.57 (Rook 1st and Queen's 1st) to 0.59 (*k*-Nearest neighbors). Median monthly housing costs are also highly spatially correlated, with values around 0.75. Percent change in per capita income from 2007 to 2008 is also highly correlated (0.528 in the Queen's 1st) indicating that counties near each other experienced economic changes from the recession in a similar fashion. Recovery was also correlated at a similar level (0.5 in the Queen's 1st).

As the dependent variable is a rate using an event (into-county moves per year's worth of county population), I also calculate the Moran's *I* statistic for county-to-county migration rate using the empirical Bayes (EB) adjustment to account for the variance

[3] Statistical inference for Moran's *I* is derived from a random permutation procedure that recalculates the statistic to create a reference distribution which is compared to the original statistic to generate a pseudo-significance level (Anselin, 2005). In this case I generated 999 permutations to ensure the statistical significance of all Moran's I values in the model.

instability of rates present in a Moran's *I* scatter plot and LISA statistic (Anselin, 2005). Even after this adjustment, the dependent variable Moran's *I* is still approximately 0.18 in all models. Although the Moran's *I* values using the *k*-Nearest neighbors 4 weights matrix are slightly higher, past research using county-level data recommends using the Queen's 1st order weights matrix (Voss et al., 2006). Moreover, the maximum likelihood estimator used to calculate the regression coefficients with a weights matrix in a spatial error regression model works best when the weights correspond to a symmetric contiguity relation (i.e. Queen's or Rook's 1st) and not for *k*-Nearest neighbors (Anselin, 2005). As later analysis reveals the necessity for such a model, I select the Queen's 1st order weights matrix to produce local indicators of spatial autocorrelation maps as well as perform spatial regression modeling.

Local spatial autocorrelation can be demonstrated by producing maps that highlight areas of high spatial autocorrelation for both high and low values on the attribute of interest. Figure 7 is a LISA cluster map that highlights both the areas of spatial clustering of high values and low values of the county-to-county migration rate. The maps in this analysis are compared to a reference distribution of 9999 permutations of the LISA statistic and are significant at below the 0.05 level, thus the highlighted counties are robust indicators of spatial clusters on the dependent variable. Looking at the map, it is evident that spatial autocorrelation exists within the states of Texas and Oklahoma as well as in north Florida, where there are high rates of county-to-county migration in the same areas. This may be due to booms in local energy economies. Moreover, counties that border northern California along with counties in the mountainous states of Colorado and Wyoming exhibit clusters of high-high internal migration. As a total, approximately 6% (193 total) of counties demonstrate high-high clustering of internal migration. On the other hand, an even greater percent of the country

(10%) including parts of the Midwest (specifically Ohio and the Detroit metro area), the Appalachian region, and the northeast megalopolis, exhibit levels of spatial clustering around low-low values, i.e. very few people are moving within the state in these regions. This may be due to limited resources (Appalachia and Detroit) or high residential stability (Northeast).

Figure 8 about here

Figure 8 shows the clustering of the local spatial autocorrelation dependency ratio. It is evident that this variable possesses a greater level of local spatial autocorrelation. What is striking is that high dependency ratios (greater proportion of very young and elderly) for neighboring counties are also clustered in Texas and Florida, but not in the counties where high in-migration clustering is present. There is also substantial clustering of high-high counties in the Plains. This could represent a causal effect, where single working age adults move within the state or into a new state to other counties for work thus increasing the dependency rate (by lowering the number in the denominator). In order to examine this relationship, a regression framework will be necessary to capture the relationship between county-to-county migration and the dependency ratio.

Analytical Framework

Traditional inferential statistics would apply an ordinary least squares (OLS) regression of the county-to-county migration rate on our set of independent variables, written as follows;

$$y_i = \beta_0 + \beta_1 x_i \ldots \beta_k x_n + \varepsilon_i \qquad (1)$$

Here, our dependent variable *y* is a function of our independent variables (x) and their linear relationship with the dependent variable (*B*) and an error term. However, this specification does not take into account the geographic location of our units nor their association with the linear function of their neighboring units. In this investigation, I initially run an OLS regression of the county-to-county migration rate on the dependency rate, the percent of county residents living in poverty, the county unemployment rate, population, percentage of wage increase (or decrease), number of counties, median monthly household costs, and USDA typology dummy variables for traditional mining and retirement counties along with five demographic county profiles groups. I then run an OLS model including the geographic Queen's 1st order weights and use model diagnostics to determine if spatial analysis is necessary. Finally, I use model fit tests to determine whether this particular research question requires a spatially lagged dependent variable (spatial lag model) or a spatial autoregressive error term (spatial error model). In spatial lag models, spatial autocorrelation is accounted through the linear relationship of the dependent variable and a spatially lagged dependent variable specified as follows;

$$Y = X\beta + pWY + \varepsilon \tag{2}$$

In this equation, the spatial autocorrelation is modeled through the function of *Y* and the spatially lagged dependent variable (*pWY*), thus variance that is spatially explained is explained through this relationship, whereas regular linear relationships are explained through the coefficients of our regular parameters. This type of model is frequently used when the spatial generative process is thought to be dependent on the variable itself (i.e., spatial diffusion of an idea)(Voss et al., 2006).

The spatial error model is specified as follows;

$$Y = X\beta + u, \qquad u = pWu + \varepsilon \tag{3}$$

In this equation, the spatial autocorrelation is captured in *u*, which is a spatially lagged error term, which implies that the errors are what are spatially correlated. This type of model is frequently used when the spatial process is hypothesized to be due to grouped responses on an unobserved value (for example, job opportunities) (Voss et al., 2006). Through this analytical framework, I will examine the constituent determinants of the internal migration to detail if there is any spatial structure to the linear relationship, how accounting for this impacts the original model parameters, and what is the likely determinant of the spatial process of internal county-to-county migration.

Model Comparison Results

The results of a regular OLS regression of county-to-county migration on the independent variables reveal a number of significant relationships. As predicted, the dependency ratio is negatively associated with internal migration, with a 1-unit increase in the dependency ratio resulting in a substantive decrease in the migration rate by nearly 1 person per thousand living in the county (coefficient = -1.03 significant at below the 0.001 level). Unemployment, intuitively, has a negative association with the dependent variable, with a 1 percent increase in the unemployment rate resulting in approximately 1 less new person per 1000. The percent in poverty is perhaps surprisingly positively associated with internal migration, with a 3 percent increase in residents living below the poverty line resulting in a 4 person increase of the internal migration rate. Notably, percent wage increase is positively associated with internal migration (a 3% increase in wages associated with 1 new person per 1000). Monthly housing is positively associated with migration, with a $100 increase in monthly housing costs associated with a 2-person increase in the in-migration rate. Furthermore, the USDA designation for mining counties is associated with an 8 person reduction in the internal migration rate while retirement counties correlate with a 5 person increase.

The divisional variables tell the story of regional migration drivers. Here, in comparison to the West South Central division (which includes Texas and Oklahoma), New England and Mid-Atlantic counties attract 17 less people per thousand, Eastern Central divisions receive less internal migrants (9 and 13 for both North and South respectively), while South Atlantic counties receive 4 less migrants per thousand. Only the states in the Mountain division receive more, attracting approximately 12 more internal migrants per thousand in comparison to the West South Central division. This regional variation mirrors the quintile map of internal migration that showed hot spots of internal migration in Texas, Oklahoma, and the Great Plains counties.

Lastly, the five cluster variables explain a distinct portion of the variance the internal migration rate. In comparison to cluster 1 (large towns), cluster 2 (small cities) and 3 (large cities) do not vary significantly, as much of their variability in internal migration may be explained by population, wages, and housing costs. Since these variables are included in the model, they have explained the variance that might have been captured in these cluster variables. However, clusters 4 and 5 (small towns and small energy towns) are both statistically significant and both receive more internal migrants (roughly 6 more per 1000) in comparison to large towns.

Table 3 about here

While these results are informative, substantively they only explain approximately 24% of the variation in internal migration. Checking model diagnostics indicates that there may be an issue with the error structure of the OLS specification. The Jarque-Bera value, which suggests non-normality of error structure, is significant at below the 0.001 level. Moreover, the Breausch-Pagan and Koenker-Bassett tests act as gauges of non-constant error variance, and since all three are significant at below the 0.001 level, this points out heteroskedasticity in the residuals and provides support for testing the model when accounting for the spatial structure of the variables. Plotting the residuals of the OLS model to a standard deviation map (seen in Figure 10) once again shows clustering of counties three standard deviations from the mean throughout Texas, Oklahoma, parts of counties near northern California, and northern Florida; however, this time the clustering is in the error term, indicating that the regular OLS model is failing to explain groups of counties in these areas. This is supported by the Moran's *I* of the residual term (Figure 11), which reveals correlation of residuals with neighbor residuals at a value of 0.15.

Figures 10 & 11 about here

Running the OLS regression again when including the Queen's 1st order spatial weights allows for testing of spatial autocorrelation in the linear model through six diagnostic tools; The Moran's *I* error value, and five unique LeGrange Multiplier test statistics that serve to inform the choice of specific spatial regression models.[4] In Table 3, the Moran's *I* of the residuals is 0.15 (previously mentioned) and the resulting z-value is fairly large (14) and is significant at below the 0.001 level, illustrating the presence of spatial autocorrelation in the model as well as model misspecification in general. Of the five Legrange Multiplier test statistics (hereafter shortened to LM), the first two (LM Lag test and Robust LM Lag test) serve to suggest the need for a spatial lag model. In this case, both of these diagnostic statistics are significant at below the 0.001 level and have values great enough to indicate a spatial lag model is necessary (120 and 23 respectively). However, the subsequent two LM test statistics (LM error test and Robust LM error test) serve to provide support for a spatial error model. As both of these tests are highly significant in the model and the test values are actually bigger than those in the previous LM-lag tests (190 and 93 respectively), it signals the need to compare the models based on model fit diagnostics.

Spatial Lagged Model and Spatial Error Model Evaluation

Model comparison in this study will focus on the post-estimation fit statistics in order to provide support to select an appropriate specification, prior to discussing the

[4] The fifth LeGrange Multiplier test, a test for the LM (SARMA) model, tests for the necessity of a higher order alternative of a model with both spatial lag and spatial error terms. But as this test statistic value was only slightly higher than the spatial error value, it will not be addressed here and is considered beyond the scope of this paper.

substantive implications of coefficients. In this case, I compare the Log Likelihood value, as well as the Akaike's Information Criterion (AIC) and Schwartz's Bayesian Information Criterion (BIC) measures, which test model fit while penalizing overly complex models (Chi and Zhu, 2008). When running both models using the Queen's 1st order weights matrix, it is clear that the spatial error model is a better fit to the data. It possesses a greater Log Likelihood value (-14247 in comparison to -14276) and lower AIC (28538 compared to 28598) and BIC (28670 compared to 28737) values as well. Moreover, both of these models possess lower AIC and BIC values as well as greater Log Likelihood values in comparison to the original OLS model, further supporting the need for a spatial model. Exploring the Moran's *I* residuals of both the spatial lag and spatial error models (approximately 0 in both) shows that both models explain practically all of the spatial dependence that was previously found in the model. Thus, model fit statistics, theory on the spatial process of grouped responses, and past studies involving county migration supports the use of the spatial error model to explain the original research questions (Voss et al., 2006.

Discussion and Results

In comparison to the original OLS results, the spatial error model not only explains the spatial autocorrelation present in the model, but also increases the impact of the dependence ratio (-1.03 to -1.07) along with increases in the impact of percent in poverty, percent unemployed, and household wages. However, the use of the spatial error model reduces the impact coefficients that were highly spatially correlated, such as divisional variables. Accounting for the spatial structure also reduces the impact of retirement counties, as these may be clustered in certain destination areas (Arizona and Florida). Yet even when controlling for spatial autocorrelation in the error term, small town and small energy town clusters are robust. Interestingly, calculating the spatial structure had very little effect on percentage increase in wages (actually increasing its positive effect by 0.03). This indicates that spatial structure may not explain the impact of increases in average wages, or affect its relationship with internal migration. Although these changes are slight, it does imply that the spatial error model is more robust, as including the spatial autoregressive error term accounts for the spatial structure of the residuals. The standard test of model specification (wald figure comparison > likelihood ratio test > LM statistic) reveals that while this model does account for spatial dependence in the error term, there is still heteroskedasticity in the model, indicating that the model is still slightly underspecified. This is a fair assessment, as there are many plausible independent variables that could explain internal migration and this model explains only 30% of the variance in the dependent variable (R^2 of 0.30). The results from this model illustrate the necessity for a spatial error model, but call for a better-specified model that accounts for more variation in the county-to-county in-migration rate.

By finding support for the spatial error model, this study helps to explain the spatial process of internal migration, attesting to the likelihood of internal migration rates being spatially grouped due to omitted economic and socio-demographic measures. Furthermore, it is clear from the spatial error model and the figures of the LISA cluster maps for the county-to-county migration rate and dependency ratio that the two variables are inversely related. This relationship can be considered even more robust now that the spatial structure is accounted for along with classical predictors of internal migration (rebounding of wages after the recession, retirement sites, metro areas, regional differences, and housing costs). Therefore, these results clearly depict the spatial process of how internal migration rates associate with population demographics; specifically, that people move to counties with greater proportions of working-age residents. Yet, this finding is attenuated when accounting for the results of clusters 4 and 5 (small towns and small energy towns). Despite high dependency ratios, these areas attract internal migrants even when accounting for recovery from the great recession. Cluster 5 (small energy boom in the Dakotas) may reveal labor migration into small towns with families that are experiencing rapid year-to-year personal income due to expansion in the energy sector. This may be explained by an industry-specific style of migration that draws people to a geographic region due to natural resources as opposed to being drawn to cities because of the greater economic possibility.

Future directions for this study would include the refinement of independent variables and further exploration of regional drivers of county-to-county migration. Ideal research conditions would allow for the examination of the attributes of movers and stayers to account for migrant selectivity, but individual-level information based on county-to-county moves is currently restricted and industry-specific county-level profits are unfeasible to include as information is frequently not disclosed by county governments and there would be reason to believe this data would be missing not at random. Gaining access to restricted individual-level data would allow for a deeper analysis of the process of internal migration. Given the robust negative relationship of the dependency ratio and internal migration, individual information on who moves from county-to-county would help to better understand this dynamic demographic process.

References:

Anselin, L. (1988). *Spatial econometrics: methods and models* (Vol. 4). Springer.

Anselin, L. (2005). Exploring spatial data with GeoDaTM: a workbook. *Urbana*.

De Jong, G. F. (1999). Choice processes in migration behavior. *Migration and restructuring in the United States*, 273-293.

Kelejian, H.H. and Prucha, I.R. (2010) Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances, *Journal of Econometrics*, 157: 53-67

Lee, E. S. (1966). A theory of migration. *Demography*, *3*(1), 47-57.

Massey, D. S., Patrikios, H., Fagbule, D. O., Olaosebikan, A., Parakoyi, D. B., Williams, A. N., ... & Theriault, G. (1990). Social structure household strategies and the cumulative causation of migration. *Population index*, *56*(1), 3-26.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*,*37*(1-2), 17-23.

Nord, M. (1998), Poor People on the Move: County-to-County Migration and the Spatial Concentration of Poverty. *Journal of Regional Science*, 38: 329–351

Ravallion, M., & Wodon, Q. (1999). Poor areas, or only poor people?. *Journal of Regional Science*, *39*(4), 689-711.

Voss, P. R., Long, D. D., Hammer, R. B., & Friedman, S. (2006). County child poverty rates in the US: a spatial regression approach. *Population Research and Policy Review*, *25*(4), 369-391.

Table Appendix

Table 1: Selected Descriptive Statistics for Dependent and Independent Variables

Measures of Central Tendency and Dispersion	Mean	Standard Deviation	Min	Max	25 ptl	50 ptl	75 ptl
Total County-to-County Migration Rate	62.58	27.87	0.00	373.67	44.38	56.69	74.41
Dependency Ratio (*100)	68.60	9.66	6.49	120.93	62.40	67.89	73.89
Percent in poverty	15.19%	6.11	0.00	46.00	10.80	14.52	18.61
Percent unemployed of people 16+	8.58%	3.72	0.00	26.80	6.12	8.31	10.66
Population (in thousands)	98.76	314.90	0.09	9840.02	11.31	25.92	66.98
Percent Wage Increase	5.42%	7.52%	-51.58%	46.38%	1.08%	4.80%	9.12%
Median Monthly Housing Costs	723.68	282.09	134.00	2405.00	544.00	652.00	832.00
Percent change in per capita personal income '07 to '08	3.78%	4.86%	-33.55%	45.57%	2.38%	3.54%	4.92%
Percent change in per capita personal income '11 to '12	6.03%	6.42%	-28.51%	74.17%	2.57%	4.73%	7.74%
Number of Counties in State	92	50	1	221	62	82	101
Mining County	0.04	0.19	0	1			
Retirement County	0.14	0.35	0	1			
Spatial descriptive statistics	Moran's *I* Q1	Moran's *I* R1	K nearest neighbors (4)				
Total County-to-County Migration Rate	0.180***	0.181***	0.186***				
Dependency Ratio (*100)	0.441***	0.446***	0.469***				
Percent in poverty	0.458***	0.458***	0.466***				
Percent unemployed of people 16+	0.570***	0.570***	0.594***				
Population (in thousands)	0.357***	0.360***	0.371***				
Percent Wage Increase	0.197***	0.196***	0.206***				
Median Monthly Housing Costs	0.738***	0.741***	0.754***				
Metro Area (2004 RUCC)	0.393***	0.396***	0.431***				
Number of Counties in State	0.920***	0.917***	0.931***				
Mining County	0.300***	0.302***	0.313***				
Retirement County	0.305***	0.309***	0.306***				
Percent change in per capita personal income '07 to '08	0.528***	0.531***	0.541***				
Percent change in per capita personal income '11 to '12	0.500***	0.504***	0.505***				
Cluster 1	0.255***	0.258***	0.263***				
Cluster 2	0.378***	0.380***	0.386***				
Cluster 3	0.290***	0.295***	0.333***				
Cluster 4	0.403***	0.407***	0.424***				
Cluster 5	0.415***	0.414***	0.437***				
	Moran's I EBRS	Moran's I EBRS	Moran's I EBRS				
Total County-to-County Migration Rate	0.180***	0.180***	0.185***				

Table 2 - Descriptive Statistics for Five K-means Cluster Variables

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Total Internal Migration rate	58.25	67.53	40.1	61.44	62.89
Dependency Ration	73.41	60.81	59.64	80.73	77.47
Percent living in poverty	16.43%	14.62%	14.13%	13.57%	11.38%
Percent of household wage changes	5.32%	4.18%	2.16%	9.31%	11.60%
Percent with college degrees	40.54%	43.31%	46.82%	45.18%	47.10%
Population (in thousands)	35.37	137.97	2460.32	11.4	8.09
Percent of population over 16 and unemployed	8.98%	9.06%	10.40%	5.89%	3.40%
Percent foreign born	3.63%	4.98%	25.03%	4.08%	2.74%
Median monthly rent	635	867	1338	566	529
Percentage per capita persona income change '07 to '08	4.30%	3.50%	3.10%	-3.20%	18.20%
Percentage per capita persona income change '11 to '12	5.70%	3.70%	1.80%	12.50%	23.30%
N	1336	1354	28	238	108
percentage of counties	42.97	43.55	0.9	9.1	3.47

Table 3: Spatial Model Comparison = Regression of County-to-County Migration Rate

	OLS Regression	OLS with Queen's 1 weights matrix	Spatial Lag Model	Spatial Error Model
Constant	103.74***	103.74***	83.67***	101.67***
	15.77	15.77	12.59	14.66
Dependency	-1.03***	-1.03***	-0.98***	-1.07***
	-14.24	-14.24	-13.84	-14.83
Percent Poverty	1.34***	1.34***	1.39***	1.57***
	12.38	12.38	13.22	14.55
Percent Unemployment over age 16	-1.08***	-1.08***	-1.13***	-1.17***
	-6.24	-6.24	-6.74	-6.53
Population (in thousands)	-0.02***	-0.02***	-0.01***	-0.01***
	-6.94	-6.94	-6.65	-6.66
Percent Household Wage Increase from '08 to '12	0.33***	0.33***	0.33***	0.36***
	5.22	5.22	5.29	5.85
Median Monthly Housing Cost	0.02***	0.02***	0.02***	0.02***
	6.90	6.90	6.87	6.55
Number of Counties	0.07***	0.07***	0.05***	0.07***
	6.64	6.64	5.10	5.10
Mining County	-8.15***	-8.15***	-7.46***	-6.45**
	-3.55	-3.55	-3.34	-2.72
Retirement County	5.198***	5.19***	4.76***	4.56***
	3.85	3.85	3.63	3.29
Divisional variables (West South Central reference)				
New England	-17.13***	-17.13***	-14.01***	-16.60**
	-4.58	-4.58	-3.85	-3.18
Mid-Atlantic	-17.25***	-17.25***	-12.84***	-16.07***
	-6.42	-6.42	-4.88	-4.34
East North Central	-9.45***	-9.45***	-6.11**	-8.70**
	-4.92	-4.92	-3.24	-3.27
West North Central	3.00	3.00	3.62*	4.33
	1.65	1.65	2.04	1.76
South Atlantic	-4.16*	-4.16*	-3.25	-2.85
	-2.38	-2.38	-1.92	-1.17
East South Central	-12.77***	-12.77***	-10.16***	-12.12***
	-6.52	-6.52	-5.31	-4.53
Mountain	11.83***	11.83***	9.24***	13.65***
	5.12	5.12	4.08	4.27
Pacific	1.00	1.00	0.69	2.13
	0.35	0.35	0.25	0.53
Cluster variables (Cluster 1 reference)				
Cluster 2	1.30	1.30	1.52	2.07
	0.97	0.97	1.17	1.57
Cluster 3	1.40	1.40	2.23	1.50
	0.20	0.20	0.33	0.22
Cluster 4	6.32***	6.32***	5.65**	4.53*
	3.51	3.51	3.23	2.46
Cluster 5	6.47*	6.47*	6.24*	7.41**
	2.48	2.48	2.46	2.68
Spatial model-specific coefficient			0.28***	0.35***
			11.28	13.94
Model Fit				
R2	0.24	0.24	0.28	0.30

Log Likelihood value	-14331	-14331	-14276	-14247
Akaike's Information Criterion (AIC)	28707	28707	28598	28538
Schartz's Bayes Information Criterion (BIC)	28840	28840	28737	28670
Test on Normality of Errors				
Multicollinearity Condition Number	50.4	50.4		
Jarque-Bera	19399***	19399***		
Heteroskedasticity of Random Coefficients				
Breusch-Pagan test	959***	959***	925***	850***
Koenker-Basset test	140***	140***		
Spatial Diagnostics				
Moran's I of model residuals		0.15	0.03	-0.01
Moran's I (error) value		14***		
Moran's I of predicted error			0.17***	0.16***
Likelihood Ratio Test for Spatial Dependence			112***	170***
Legrange Multiplier (lag)		120***		
Robust Legrange Multiplier (lag)		24***		
Legrange Multiplier (error)		190***		
Robust Legrange Multiplier (error)		94***		
Legrange Multiplier (SARMA)		214***		
All models N of 3109 p values >.05 = * >.01 = ** >.001=***, T statistics reported below coefficients				

Figure Appendix

Figure 1: Histogram of County-to-County Migration Rate

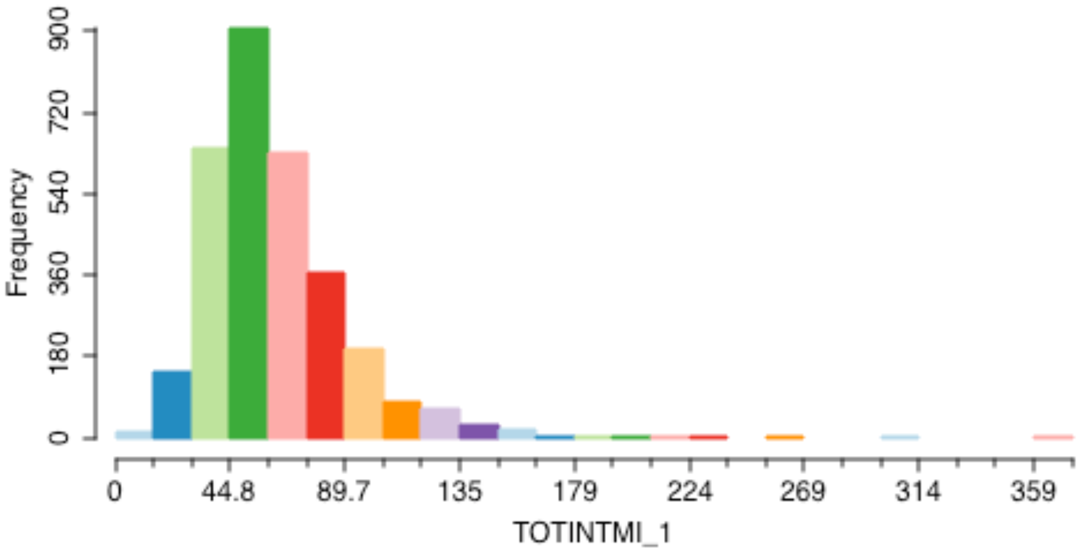


Figure 2: Quintile Map of County-to-County Migration Rate

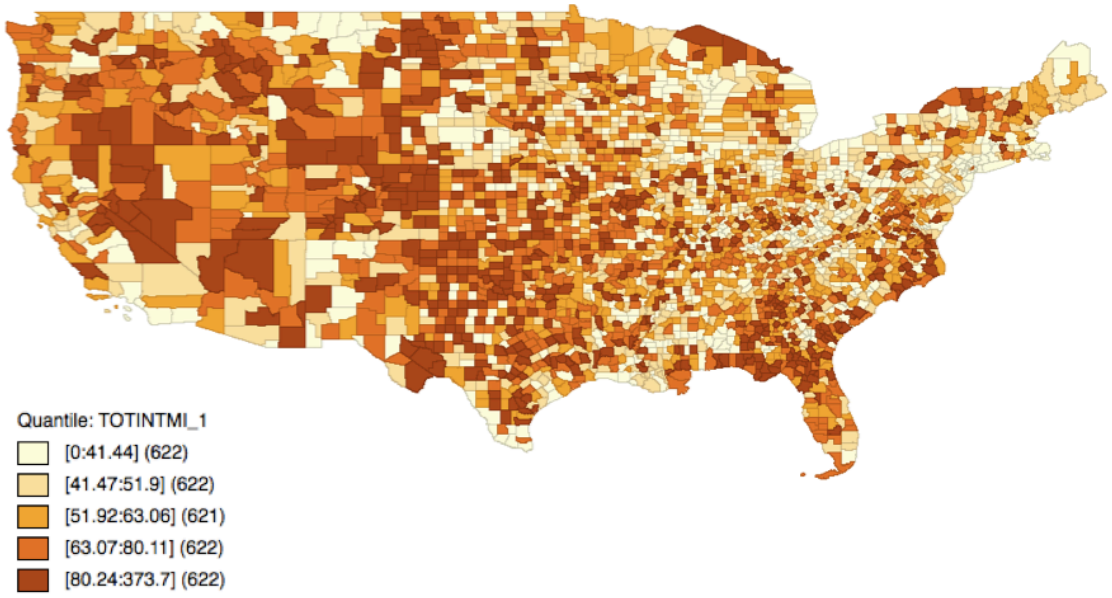


Figure 3: Histogram of the County-level Dependency Ratio

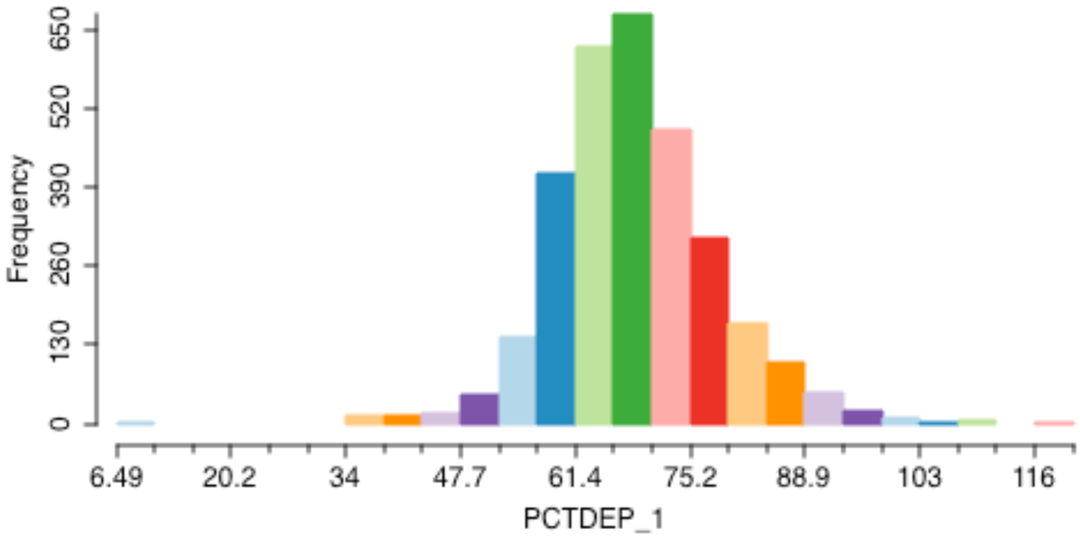


Figure 4: Quintile Map of the Dependency Ratio

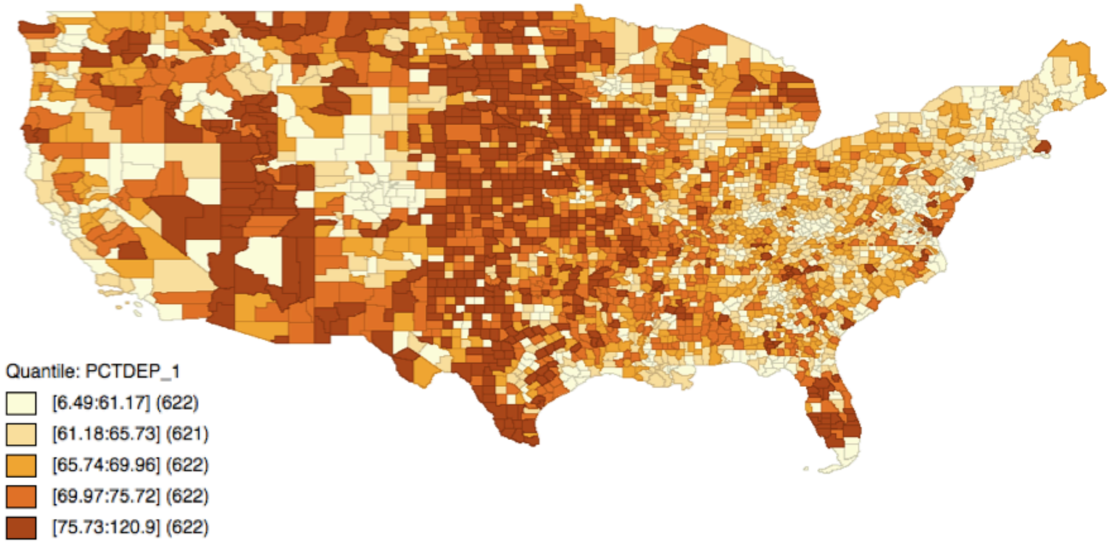
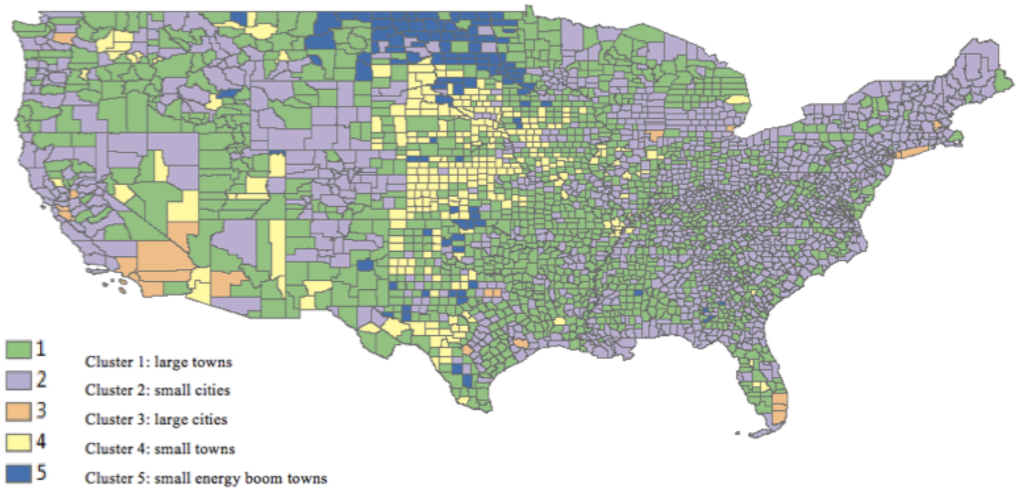
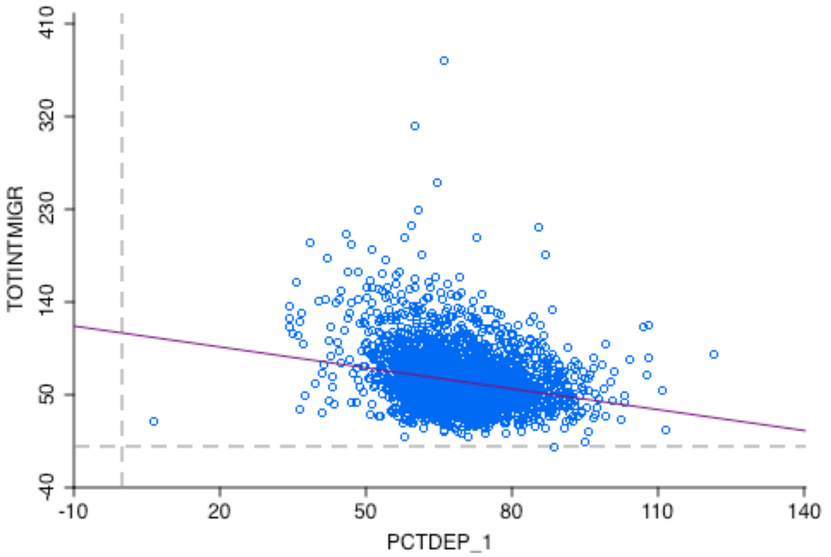


Figure 5: Cluster map of five K-means sorted variable groups as demographic county profiles



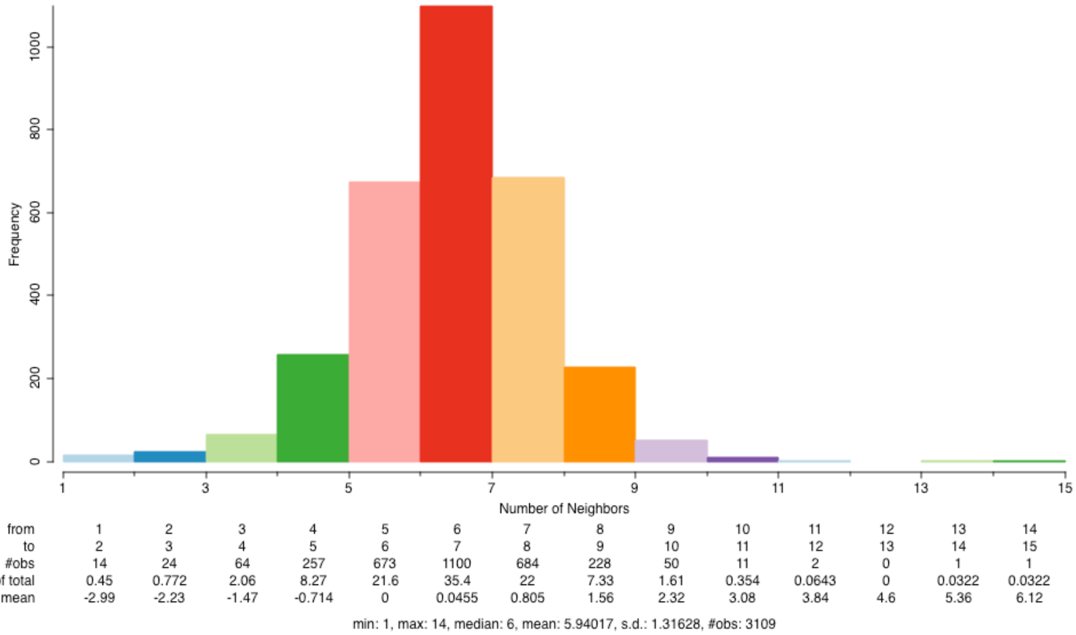
🟩 1 Cluster 1: large towns
2 Cluster 2: small cities
3 Cluster 3: large cities
4 Cluster 4: small towns
5 Cluster 5: small energy boom towns

Figure 6: Bivariate scatterplot of County-to-County Migration and the Dependency Ratio



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
3109	0.0536	108	3.49	31.1	0	-0.668	0.0504	-13.3	4.09e-39

Figure 7: A connectivity histogram of the Queen's 1st order weights matrix



from	1	2	3	4	5	6	7	8	9	10	11	12	13	14
to	2	3	4	5	6	7	8	9	10	11	12	13	14	15
#obs	14	24	64	257	673	1100	684	228	50	11	2	0	1	1
% of total	0.45	0.772	2.06	8.27	21.6	35.4	22	7.33	1.61	0.354	0.0643	0	0.0322	0.0322
sd from mean	-2.99	-2.23	-1.47	-0.714	0	0.0455	0.805	1.56	2.32	3.08	3.84	4.6	5.36	6.12

min: 1, max: 14, median: 6, mean: 5.94017, s.d.: 1.31628, #obs: 3109

Figure 8: Local Indicator of Spatial Autocorrelation map for County-to-County Migration Rate (calculated at 9999 permutations)

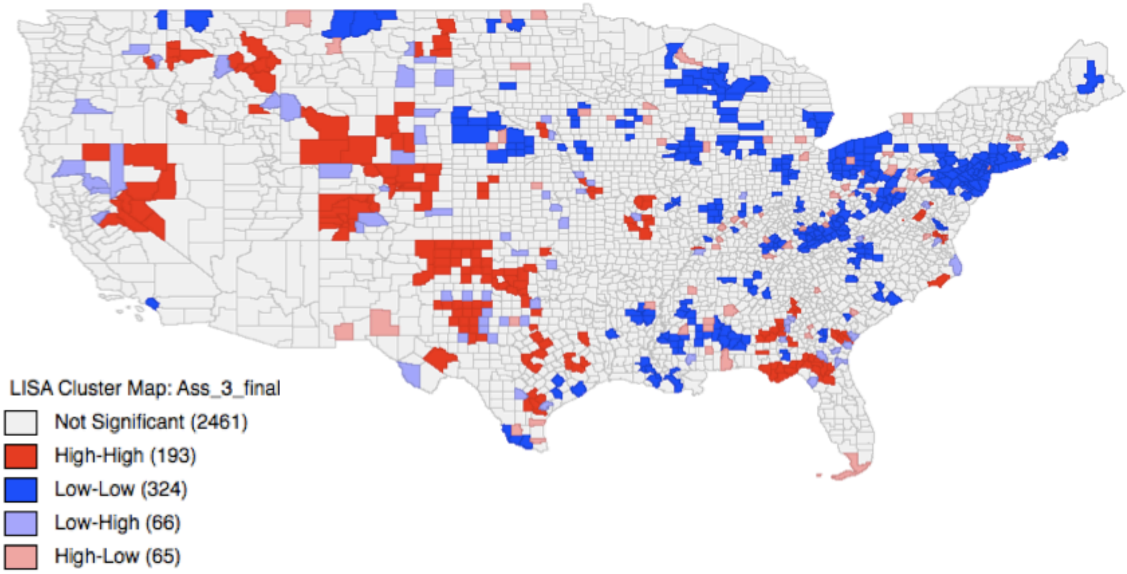


Figure 9: Local Indicator of Spatial Autocorrelation map for the Dependency ratio (calculated at 9999 permutations)

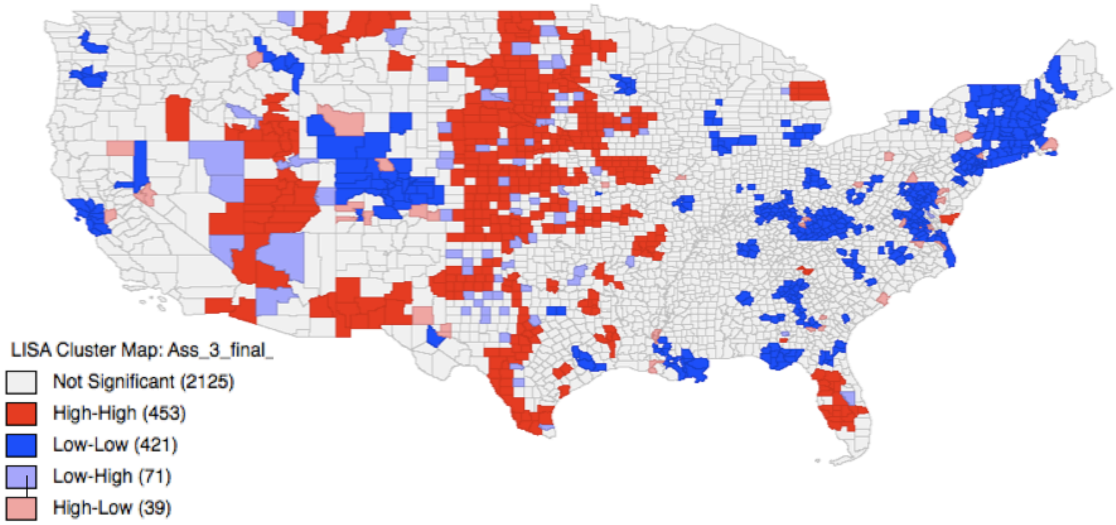


Figure 10: Standard Deviations Map of the Ordinary Least Squares Residuals

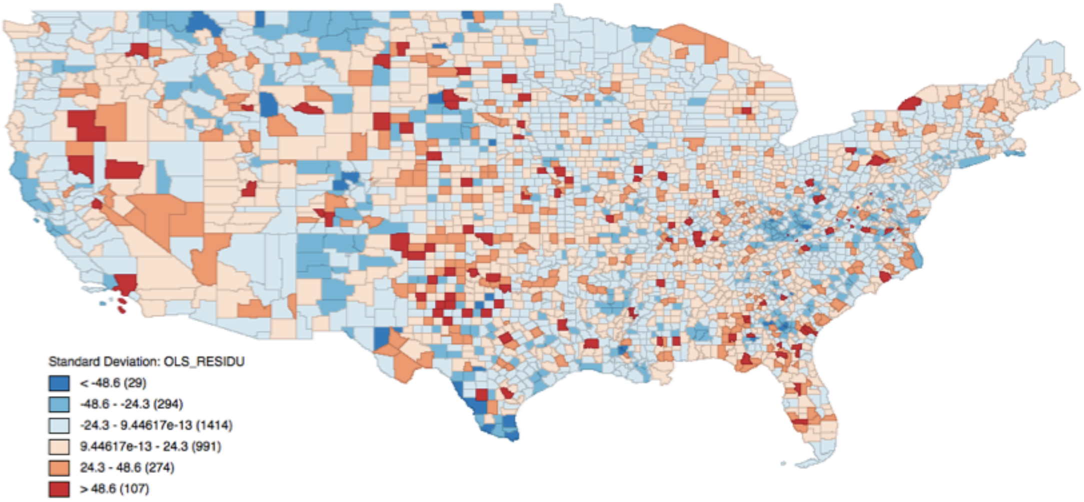
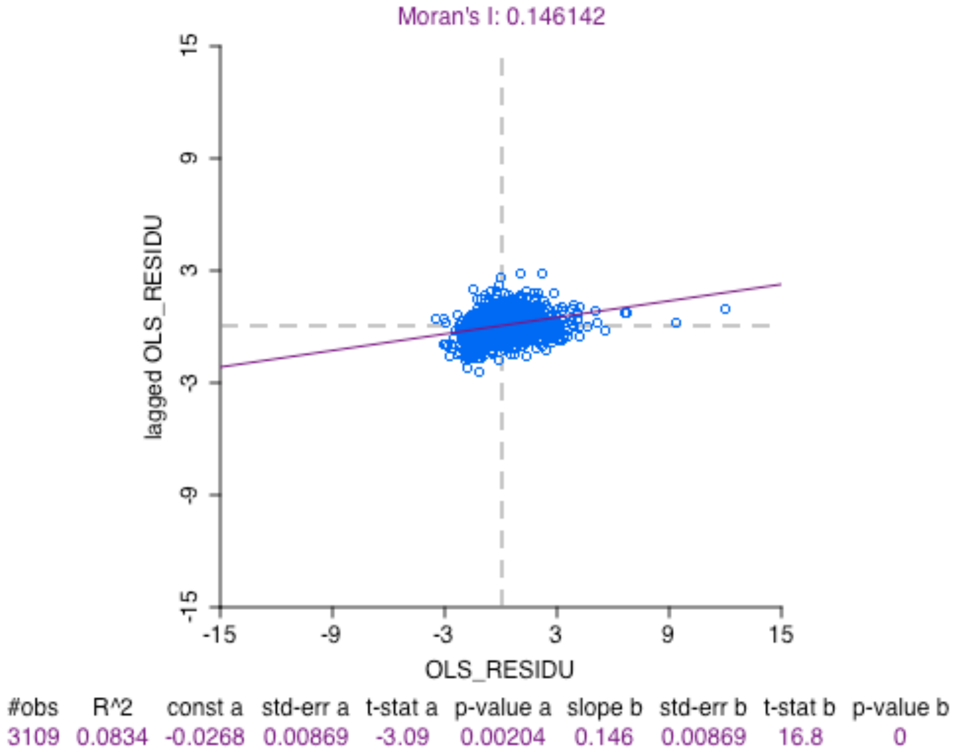
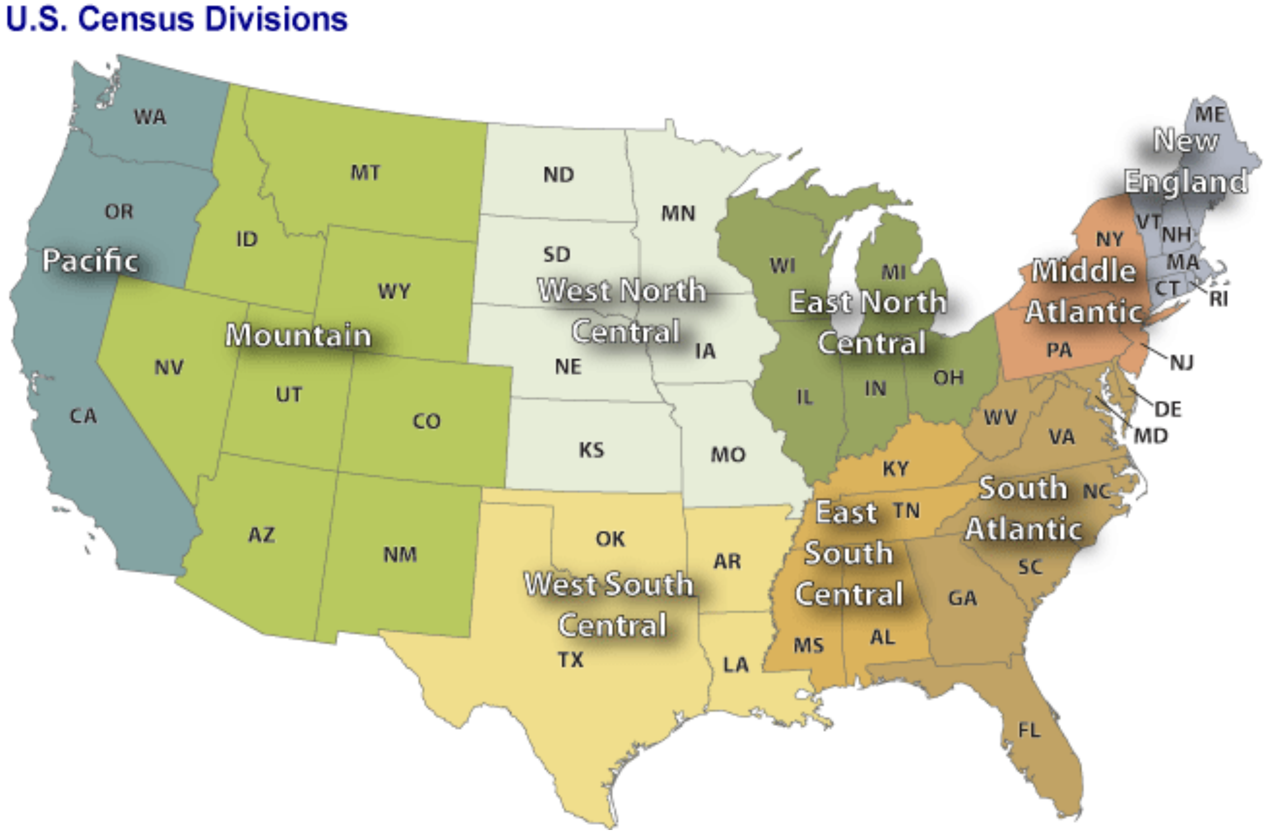


Figure 11: A Moran's *I* scatter plot of the Ordinary Least Squares Residuals



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
3109	0.0834	-0.0268	0.00869	-3.09	0.00204	0.146	0.00869	16.8	0

U.S. Census Division Distribution – image provided by U.S. Census Bureau


U.S. Census Divisions
WA
OR
Pacific
CA
MT
ID
WY
NV
Mountain
UT
CO
AZ
NM
ND
SD
MN
West North
Central
NE
IA
KS
MO
WI
MI
East North
Central
IL
IN
OH
OK
AR
West South
Central
TX
LA
KY
East
TN
South
Central
MS
AL
ME
New
England
VT
NH
NY
MA
CT
RI
Middle
Atlantic
PA
NJ
DE
MD
WV
VA
South
NC
Atlantic
SC
GA
FL