2005

Push and pull: Using regression models to study the influence of economic variables on net migration in Iowa counties

Andrea White
University of Northern Iowa

Push and pull: using regression models to study the influence of economic variables on net migration in Iowa counties

Andrea White
Senior Thesis: Presidential Scholars Program
May 4, 2005

Abstract

Migration is an intensely personal decision, but mathematical models are useful for quantifying the larger, economic aspects of it. The goal of this research is to use spatial and multiple regression models to study the influence of economic variables on net migration rates in Iowa counties. To achieve this data for many variables was collected from several sources and centered on the year 2000. S-plus software was used to create neighborhood structures, run spatial correlations and regressions, and run multiple regressions and residual diagnostics. The results showed that it is possible to develop a good regression model of migration using net migration as the dependent variable along with various economic covariates. Results also emphasized the rural nature of Iowa, as outliers were often the larger and more urban counties. Two counties in particular, Dallas and Woodbury, were extreme cases for the state of Iowa. This research shows that despite migration's subjective nature, regression models are applicable to the study of migration and can lead to a better understanding of why migration occurs.

Table of Contents

List of Figures

List of Tables

Introduction

Migration is a vital measure of society. It is unique in that unlike birth and death, which we lack control over, we maintain the power to move where we want. Thus studying migration is equally important, if not more important, than studying other vital life statistics. However, there are many approaches for studying migration and many ways to attempt to quantify this highly personal subject.

As Cadwallader (1992) explains, there are several different approaches to studying migration. The micro approach is concerned with the individual, and the psychological decision making process of migration. The macro approach looks beyond this to aggregate migration behavior, including characteristics of the socio-economic and physical environment. There are also three different schools of thought on why migration occurs. The institutional approach emphasizes the effects of institutions such as governments, real estate companies, etc. The behavioral approach looks at the process and decision-making involved rather than the pattern of migration. Finally, the neoclassical approach suggests that labor moves in response to interregional wage differences. In this view workers are assumed to maximize income and there are no barriers to labor mobility.

In reality all three of these approaches have validity. It is no doubt a combination of them that truly drives migration. Speculating on what causes migration is easy; trying to model it mathematically is a challenge. Ultimately migration is a subjective choice made on a personal or family level. Attempting to model personal choice with numbers and equations may not be an effective way of studying migration. However, there are many ways to measure variables such as migration rates, employment levels, income,

housing prices, and other economic variables. Thus the question is can a model be developed for a specific time and place that explains net migration in terms of economic variables?

Similar studies

There are many previous studies that have attempted to model migration. One example comes from Chen and Coulson (2002), who looked at the determinates of urban migration in Chinese cities. They used a regression model with the migration rate as the dependent variable and several independent variables, including gross city product, per capita gross domestic product, salary, employment rate, businesses, employment in second and third sectors, foreign direct investment, housing investment, public transport, and fiscal expenditures. Their findings were that the structure of the city's economy is what attracts migrants; in particular, cities with high ratios of employment in the manufacturing and service sectors experienced higher growth rates from migration.

This model is suggestive of the rural to urban trend which has been occurring over the last century. Cities that are growing are the ones with increasing manufacturing and service sectors, which occur in the presence of increasing urbanization. Interestingly, a study in Washington found that when new jobs are added they are usually obtained by in-migrants, and not prior county residents. An input-output model showed that 95-98 percent of the labor force change consisted of new migrants (Yeo & Holland, 2004).

Another example of a regression model comes from England (Fotheringham, Rees, Champion, Kalogirou, & Tremayne, 2004). This was a comprehensive study that reviewed data for 139 possible determinates of out-migration for 98 areas. This data was also calibrated for 14 different population groups. Examples of their findings include:

there is a greater pull out of places near large urban centers; out-migration rates increase as the proportion of non-white population in an area increases; and out-migration is higher from areas having higher incomes and from areas experiencing large volumes of employment. The general trend was that areas that were economically deprived had low out-migration rates and areas that were economically prosperous had high out-migration rates. The authors point out that it is difficult to model the true relationship between out-migration rates and origin economic variables because it is difficult to separate the desire to move from the ability to move.

Studying migration in the United States also has this difficulty. Many studies point to certain variables as being indicative of high in or out-migration, but it is impossible to tell where the variable's influence stops and personal choice starts. One study by Anjomani (2002) used a simultaneous equation model of interstate migration. The author found that neither the growth of employment nor the growth of income in the destination location were directly important determinates in migration flow. In general, states with lower income growth and higher unemployment produced out-migrants, and states with lower crime rates, lower population densities, and high population growths attracted in-migrants. Another U.S. study by Shelly and Koven (1993) used a multiple regression analysis and found that a composite of ecological, quality of life, and economic variables were most significant in predicting rates of net state migration.

Studying state to state migration in the U.S. helps paint the big picture, but there are certainly interesting dynamics occurring within states. One author (Vias, 2001) looked at variations in county migration rates and classified counties based on a variety of socioeconomic and geographic characteristics. Major findings were that large, non-

metropolitan counties that were close to metropolitan counties had exceptionally large turnover in population, and remote agricultural counties in the central U.S. have little variability in gross migration rates. Rappaport (2004) found that population flows to an area are persistent, and that local areas that are growing rapidly tend to continue to do so.

These studies suggest that there are three basic paths for a county, particularly one in the Midwest. The county can likely be classified as already urban, as becoming urban, or as rural. This seems to be a good classification scheme for Iowa counties. Iowa is a state with a relatively stable population, without major increases or decreases in migration over time. However, there have been several observable trends in recent years. Iowa has seen gains in younger ages and losses in older ages due to migration. There has also been in-migration for people educated at or below the high school level, and out-migration for people with a college education. Additionally, higher incomes tend to lead to more out-migration (SETA, 2004).

While it is interesting to view these trends at a state level, is it possible to model migration trends at a county level, and in particular, build a model showing the importance of various economic variables? Do Iowa counties exhibit the stereotypical rural to urban trend, or is there something else going on? What counties are thriving, and what counties are lagging behind? Answering these questions is the goal of the present research.

Objectives

The basic objective of the project was to develop a regression model that explained net migration in Iowa counties. It was decided to run two types of regression, spatial and multiple, and compare the models. The goal was to create a model with a

small number of variables but a high percentage of the variability of net migration explained. A final objective was to discern migration patterns in Iowa and try to understand outlier counties: those counties doing extremely well or extremely poorly.

Data Collection

Data was collected from a variety of sources. The most important data for the study was net migration rates. This data was obtained from the Internal Revenue Service. The data was received as in and out-migration by county, and subdivided within each county to show what other county or country people were moving from or to. The total in and out-migration rates for each county were combined to form the net migration rate. Since migration for one year may not be representative of the trend of an area, net migration rates were aggregated for a five year period from 1998-2002. The sum migration for these five years was used as the net migration rate.

Several variables were collected from the Iowa Department of Workforce Development, particularly the size of the civilian labor force as well as its breakdown into employment and unemployment. These variables were also aggregated, this time for the years 1999-2001, and an average was calculated. Other than migration and employment, all other variables were only for a single year, most often the year 2000.

The Office of Social & Economic Trend Analysis (SETA) and the U.S. Census Bureau were the data sources for the remaining variables. There was a wide range of variables considered, including average wage, number of jobs, per capita income, number of firms, housing units, median housing value, renter units, families in poverty, public school enrollment, crime rates, sales per capita, total population, median age, rural non-

farm population, urban population, farm population, total farm acres, and farm acre change from 1997-2002. The breakdown of the population by race was also included.

There were several initial issues with the data that had to be considered throughout the study. Foremost was the summing of five years of migration. This was important to avoid using one year that may have been misrepresentative of an area's true trend, but it may have resulted in distorted results. Also, the other variables were not able to be summed in this way. Most data was only available for the year 2000, thus an attempt was made to center all variables and migration data on the year 2000 to mitigate any effects of data differences.

Statistics

It was important to gain a feel for net migration in Iowa before running any models. Thus the first step after data collection was viewing the statistics of net migration. There were 79 of 99 counties, or almost 80 percent, that lost population from 1998-2002. Also, 74 of the 99 counties, or almost 75 percent, had a total change of less than 1000 during this time period. This suggests that while most counties are losing population, the overall change is not large. The average county migration during this period was -654.

Looking at the gainers and losers shows some interesting trends. Figure 1 shows that there are only four counties with gains over 1000, and there are 21 counties with losses of more than 1000. Further, the big gainers tend to be in the central part of the state and the big losers tend to be in the northern and far eastern parts of the state. Looking at an overlay of incorporated areas (Fig. 2) shows that many of the counties with significant changes have large urban areas within or near them.

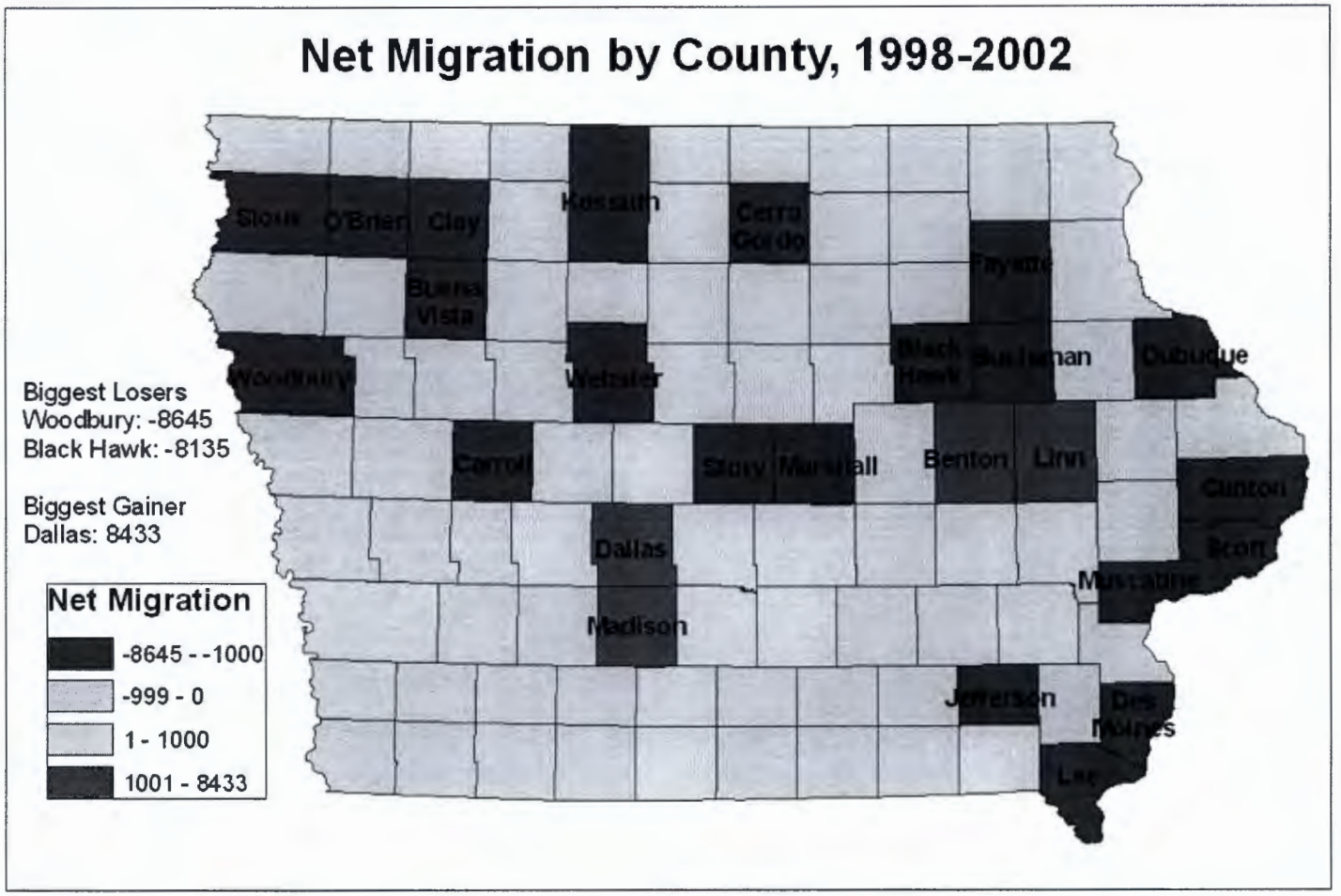

Figure 1: Net Migration by County, 1998-2002. There are only four counties that gained more than 1000 people, while there are 21 that lost over 1000 people.

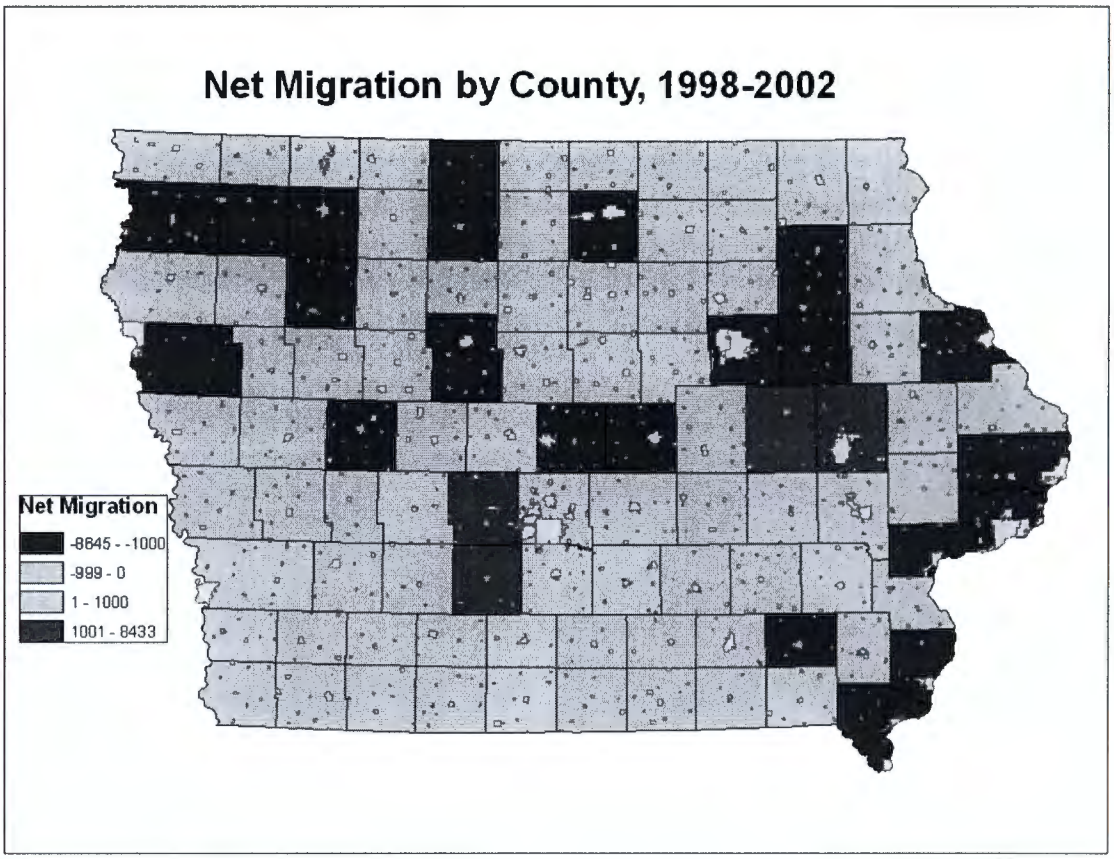

Figure 2: Net Migration by County, 1998-2002, with incorporated areas. Many of the counties with large population changes contained large urban areas.

Data Analysis

The first step in data analysis was checking for spatial correlation of net migration. This involves the development of a neighborhood structure, which gives weights to counties based on their position with regard to each other. Two structures were developed, one which included all counties touching the target county, and one which only included counties that shared a border with the target county, and not those that met at a corner (Fig. 3 & Fig. 4). The point of spatial correlation is to see if there are spatial trends in the data: for example, whether high migration rates tend to be near other high migration rates. This is measured by a p-value. A large p-value, over .15, means that there is no significant spatial correlation, while a small p-value, less than .01, means that there is very significant spatial correlation.

The Moran's I statistic was used for this spatial correlation. It was run twice, first with the neighborhood structure including corner counties and then with the neighborhood structure excluding corner counties. The former gave a p-value of .06 (moderately to strongly significant) while the latter gave a p-value of .18 (just out of the significant range). This meant that there was significant spatial correlation of net migration rates based on the first neighborhood structure.

Since there was significant spatial correlation, a spatial regression model was run. This looks at variables not only in consideration with the dependent variable of net migration, but also with the neighborhood structure. Since so many variables were being considered, the first step was running individual regressions for each variable. This allowed for the elimination of variables with very high p-values. Some surprising

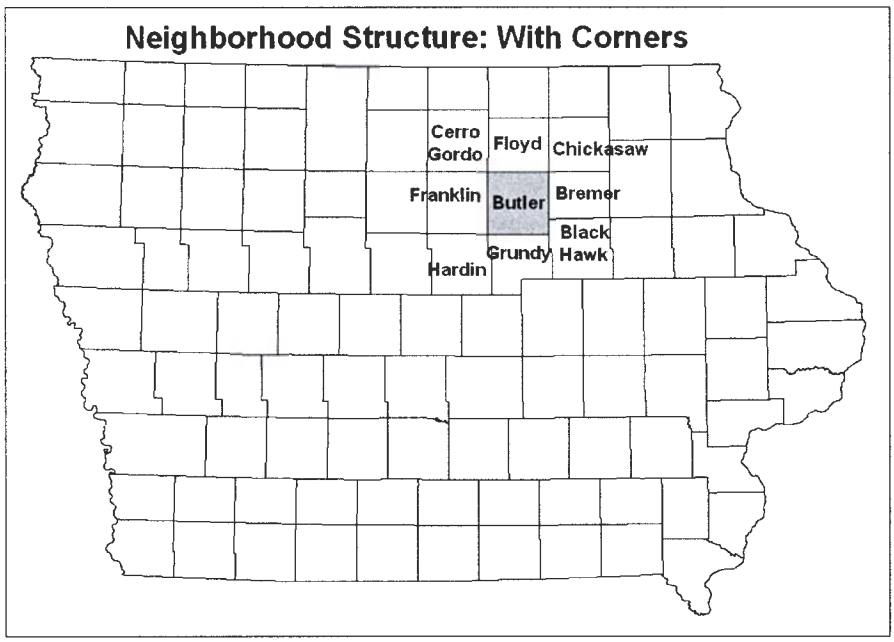

Figure 3: Neighborhood structure that includes bordering counties and counties touching only at a corner.

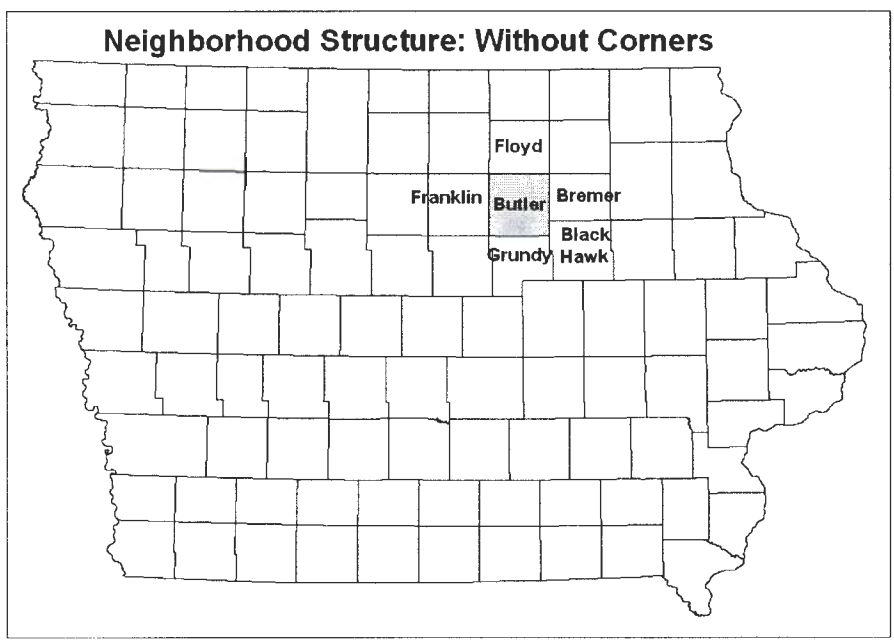

Figure 4: Neighborhood structure that includes only bordering counties.

variables included median housing value and per capita income, both of which proved to be highly insignificant in determining net migration in this model.

After an initial look at the variables, different combinations of variables were run with net migration as the dependent variable. The goal was to find the variables that were repeatedly significant, and build a model with them. S-plus software was used for all regressions, and it proved to be a difficult program to run spatial regressions in. There are three types of spatial regression possible in S-plus. The first is the Conditionally Auto-Regressive Model (CAR), which models responses given only data at neighboring counties. The second is a Simultaneously Auto-Regressive Model (SAR), which models all responses simultaneously, but incorporates the neighborhood structure. The third is a Moving Average Model (MA), which produces average values based upon neighborhood structures. The original regressions were run with the SAR model, but at times the model simply would not run and another model had to be substituted. A quality check was performed by running a set of variables with all three models, and it was found that they gave nearly identical results. Therefore, switching between models for various regressions was not considered an issue.

The final spatial regression model started with 14 independent variables, determined by their significance in earlier models. This was narrowed down to five variables by running the model, picking off the variable with the highest p-value, and rerunning the model. Ultimately all p-values were less than .01, meaning all remaining variables were strongly significant. The five variables were civilian labor force, civilian employment, civilian unemployment, housing units, and urban population. A problem observed here and in later models was that of multi-colinearity. The variables of

employment and unemployment are highly correlated, and thus can have a large influence on the outcome of the regression and whether or not it is truly accurate. This may explain their slopes, as civilian labor force had a slope of -75.3 and civilian employment had a slope of 75.3.

Following the development of a spatial regression model, multiple regression was attempted. The first multiple regression model used those variables deemed significant by the spatial regression. This model yielded an r^2 value of .59, meaning the model explained 59 percent of the variability of net migration. An r^2 value closer to one is preferable, so another multiple regression was modeled.

The second model began with a wide net of variables, similar to the beginning of the spatial regression model. Ten variables were left as significant through the narrowing of this model (table 1). Four of the five variables from the spatial regression were present, along with housing variables, sales per capita, and percents white and black. This model had an r^2 of .76. This was an improvement on the original model, but still low for the high number of variables.

In order to refine the regression, residual diagnostics were performed. These are plots to show how well the model fits the data. Two of them, the Fit vs. Residuals plot and Normal QQ plot, can be used to identify outliers. The Cook's Distance plot identifies the data values that have the most influence on the model. These three plots were run and it was specified that the five most extreme counties be identified. The results were three each lesser outliers and influential, and two counties that were both outliers and influential (Fig. 5).

Variable	Slope	p-value
Civilian labor force	-99.6383	0.0003
Civilian employment	99.7593	0.0000
Civilian unemployment	90.3913	0.0000
Unemployment percent	609.9658	0.0000
Housing units	0.3006	0.0033
Median housing value	0.0284	0.0004
Renter units	-0.6152	0.0000
Sales per capita	-0.1544	0.0032
Percent white	145.3119	0.0025
Percent black	500.0631	0.0035

Table 1: Variables included in multiple regression model, before diagnostics.

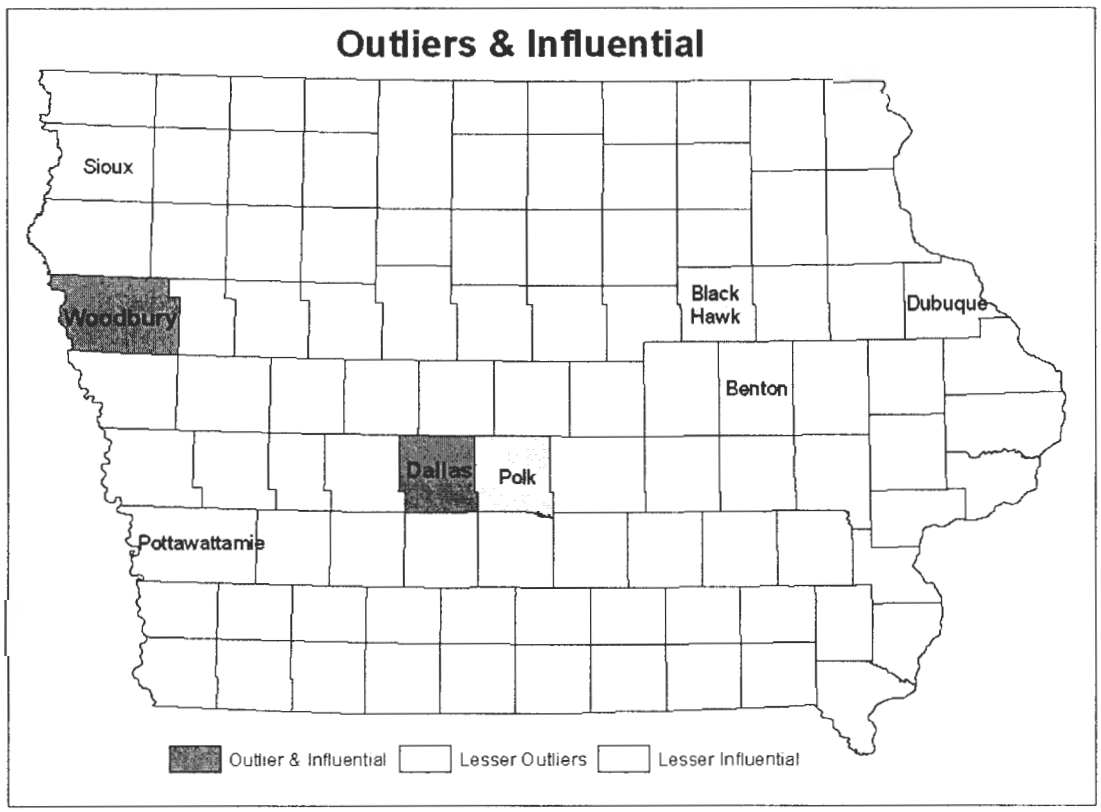

Figure 5: Outliers and influential counties from first regression model. Dallas and Woodbury counties are both outliers and influential.

The lesser outliers identified were Benton, Sioux, and Pottawattamie Counties. An examination of each counties' variables revealed the likely reasons for being outliers. Benton County had the second highest in-migration rate as well as a high median housing value. Sioux County had a high median housing value considering its high level of out-migration. Pottawattamie County was in the top ten for in-migration, and also had a high sales per capita.

The lesser influential counties identified were Black Hawk, Dubuque, and Polk. Black Hawk County had the second highest out-migration rate for the state. Dubuque County had the fourth highest out-migration rate as well as a high median housing value. Polk County has the largest population in Iowa, and with that comes high numbers in many of the variables considered. Polk County also had a very high sales per capita.

Two counties were identified as both outliers and influential, meaning they deviate from the majority of the data points and are exerting a strong influence on the model. The first is Dallas County. This county had the highest in-migration rate for the 5 year period, far surpassing the next highest county. Dallas is an average size county, but its location directly west of Polk County and the Des Moines metropolitan area result in it having low unemployment and high median housing values. The second extreme county is Woodbury. This county experienced the highest out-migration rate for the five year period. Other significant deviant variables in this county included a large number of housing and renter units and a low percent white. It was decided to run a new regression model with Dallas and Woodbury Counties removed, in the hope of fitting a better model to the state as a whole.

The new regression model followed the pattern of previous ones, starting with many variables and gradually narrowing down to significant ones. The model started with 21 variables and had an r^2 of .88. This is a satisfying r^2, but too many variables for a good model. The variables were eventually narrowed down to six strongly significant variables, and the r^2 for this model was .82. This means the removal of 15 variables from the model resulted in a loss of only six percent of the explanation of variability of the model. The variables in this model were civilian labor force, civilian employment, unemployment percent, public school enrollment, median housing value, and percent white.

Residual diagnostics were performed on this new model. This time three counties were identified as outliers and influential: Benton, Johnson, and Story (Fig. 6). The reasons for these extremes were examined. Benton County had the same issues that made it an outlier in the previous model, namely its high in-migration rate and high median housing value. Johnson County had the highest median housing value in Iowa, as well as a low percent white. Story County also had a high median housing value and low percent white. Interestingly, Johnson County gained population during the time period and Story County lost population.

The last regression model was considered the best model from this study for representing net migration in Iowa counties. It had a high r^2 value with a low number of variables. Outliers and influential observations could continue to be removed from the model, but its accuracy for the state as a whole would decrease. Ultimately, a good model was developed with the removal of only two counties.

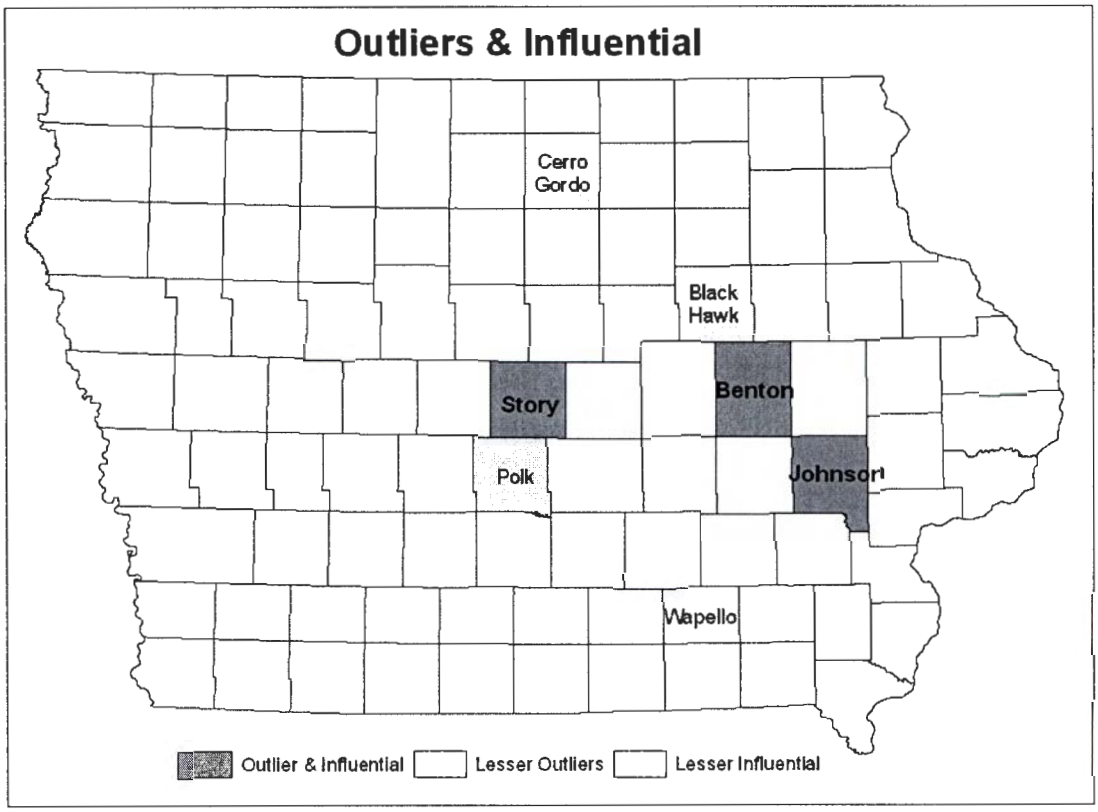

Figure 6: Outliers and influential counties from the second regression model. Story, Benton, and Johnson Counties are both outliers and influential.

Discussion

The final regression model seemed to be a good representation of important variables for net migration rates in Iowa counties. However, many issues arose during and after the data analysis portion of the project. The first issue is that a final step to the project was missing. A prediction model would have been an appropriate way to test the model built, but there were difficulties in running one. While net migration was easily obtainable for other years, the various covariates were not. Two methods of prediction were considered: either running the model on the same years for counties from another state in the Midwest, or running the model on Iowa counties for a different time period. Neither of these options proved successful as the same variables used in the model were not available for different areas or different years.

There were further issues with the data that came to light after analysis was complete. The concept of multi-colinearity was not considered prior to analysis. Combining or removing extremely similar variables may have improved the accuracy of the model. Also, absolute numbers were used for all variables except racial percentages and unemployment percent. A truer representation of Iowa may have been obtained by dividing all variables into population, since Iowa is a state with a few large counties and a lot of small counties. In addition, since the data was divided by county, moves within a county were not represented. There is much debate over when a move equals migration, and for the context of this project a move was considered migration if it crosses county boundaries.

There were some surprises from this regression analysis. While farm, urban, and rural non-farm populations were considered as variables, none of them were significant in

the final models. Farm acres and farm acre change were equally insignificant. This seems to deviate from the idea of rural to urban migration, since the population differences were so insignificant. There was also a surprise in that the percent white repeatedly showed up as significant. There was a positive correlation between percent white and net migration. The explanation for this may be that in Iowa, lower percent white populations are usually found in higher population counties, and those are the counties that are experiencing the most changes in net migration.

Residual diagnostics revealed that two counties in particular do not fit the model developed. Dallas and Woodbury counties are extreme outliers and very influential when included with all other Iowa counties. This was not surprising, as they were the highest counties for in and out-migration, but the reasons behind those rates were interesting. Outliers and influential counties tended to be the larger population counties. In Iowa, ten of 99 counties have a population over 50,000 and 63 of 99 counties have a population under 20,000. Thus it makes sense that larger counties will repeatedly be outliers. It would be interesting to run models on specific counties, grouped by population.

Conclusions & Future Directions

This project attempted to build a regression model to show the impact of economic variables on net migration in Iowa counties. Factors influencing net migration are very complex. It was impossible to narrow the variables to one or two, but it was possible to develop a good fitting model with six variables. This suggests that regression models are useful tools for studying migration.

Dallas and Woodbury counties are extreme cases for the state of Iowa from 1998-2002. Dallas showed extreme growth while Woodbury showed extreme decline. In most

cases the outliers and influential counties were higher population counties, which makes sense given Iowa's population distribution. However, there was no evidence for a strong rural to urban trend. Many of the highly urbanized counties lost large amounts of population during the study period while many of the rural counties stayed relatively constant.

Multiple regression worked better than expected to model net migration. The data used seemed incomplete, as it was not always from the same year. Also, migration is a very subjective decision. It was unclear at the beginning of analysis if it would be possible to obtain a high r^2 value, as it was impossible to input personal reasons and feelings into the model as a variable. However, the regression model did produce respectable results using only measurable variables.

There are many possible ways to continue this project. The first step would be to run a prediction model using the regression equations developed. It would also be pertinent to redo the models by trying to eliminate the very highly correlated variables and also by standardizing the variables with population. This could reveal whether the model still holds up or changes due to the different variables. It would be interesting to expand the models to include other Midwest states, and to see if they are experiencing similar net migration patterns to Iowa. It would also be interesting to distinguish between intra-state migration and inter-state migration, and see if there is a difference between the destinations of people moving from another Iowa county and people moving into the state.

Migration is an intensely subjective decision, and attempting to model it mathematically can be a challenge. However, people do move for more general reasons,

such as finding a better job, better house, etc., as well as specific personal reasons. Regression models are an appropriate way to study the influence of economic variables on migration. While these models can never truly explain migration, they can give some insight into its motivations and lead to a better understanding of why migration occurs when and where it does.

Sources Used

Anjomani, A. (2002). Regional growth and interstate migration. *Socio-Economic Planning Sciences* 36, 239-265. Retrieved February 18, 2005, from Science Direct database.

Cadwallader, M. (1992). Migration and residential mobility: macro and micro approaches. Madison: The University of Wisconsin Press.

Chen, A. & Coulson, N.E. (2002). Determinants of urban migration: evidence from Chinese cities. *Urban Studies* 39(12), 2189-2197.

Fotheringham, A. S., Rees, P., Champion, T., Kalogirou, S., Tremayne, A.R. (2004). The development of a migration model for England and Wales: overview and modeling out-migration. Environment and Planning A 36, 1633-1672.

Rappaport, J. (2004). Why are population flows so persistent? *Journal of Urban Economics* 56, 554-580. Retrieved February 21, 2005, from Science Direct database.

SETA. (2004). Characteristics of Iowa's "Movers" and "Non-movers", 1995-2000. Retrieved October 23, 2004, from http://www.seta.iastate.edu/library.aspx?lmode=1&ltopic=4

Shelley, M. C., & Koven, S. G. (1993). Interstate migration: a test of competing interpretations. *Policy Studies Journal* 21(2), 243-262. Retrieved March 3, 2005, from Expanded Academic ASAP.

Vias, A. C. (2001). Reevaluating the relationship between in-, out-, and net migration for nonmetropolitan counties: an update on Beale's U-shaped curve. *Geographical Analysis* 33(3), 228-246.

Yeo, J. & Holland, D. (2004). Economic growth in Washington: an examination of migration response and a test of model accuracy. *International Regional Science Review* 27, 205-237.