University of Nevada, Reno

Classification and Statistical Analysis of Employment Growth in United States Counties

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Mathematics

by

Benjamin H. Claassen

Deena R. Schmidt, Ph.D. \Thesis Advisor

May 2019

University of Nevada, Reno

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

BENJAMIN H. CLAASSEN

entitled

Classification and Statistical Analysis of Employment Growth in United States Counties

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Deena R. Schmidt, Ph.D. Advisor

Ilya Zaliapin, Ph.D., Committee Member

Paul Hurtado, Ph.D., Committee Member

Mark Nichols, Ph.D., Graduate School Representative

David Zeh, Ph.D., Dean, Graduate School

May 2019

Abstract

Employment growth is an important economic indicator. It is used to judge the state of the economy and drive major policy decisions. Employment growth is typically examined by analyzing company characteristics and policy initiatives. In this thesis, we engage other types of data at the county level in pursuit of explaining this growth. Specifically, we investigate how social, economic, and demographic information can explain changes in levels of employment. Principal component analysis is employed across ten county-level data sets from the United States Census to describe trends in these variables. Linear regression is used to examine connections between principal component trends and growth in economy-wide employment, labor force participation, and employment in four main economic sectors. Clear structure is observed amongst the principal components, showing strong differences across regional and urban/rural divides. Regression analysis finds significant relationships with the employment and labor force growth rates investigated. These connections indicate potential new avenues of research for policy initiatives aimed at promoting economic growth.

Acknowledgements

I want, first and foremost, to thank my advisor, Dr. Deena Schmidt. This would have not been possible without your generous and unfailing support. Thank you for your phenomenal guidance and insight, and for helping keep me focused.

I also must thank Dr. Ilya Zaliapin. Your excellent classes have been fundamental to all of my work, and your constant willingness to discuss possible problems is deeply appreciated. Thank you too to Dr. Paul Hurtado, for always keeping an eye on the big picture and for thoughtful and incisive feedback. And a particular thank you to Dr. Mark Nichols, for helping me find a project that has been both fascinating and helped me see my own future more clearly.

Finally, I am only here today because of my undergraduate advisors, Dr. Anna Panorska and, again Dr. Mark Nichols. Thank you both very much for helping me to get on this path.

Contents

List of Figures

Chapter 1

Introduction

1.1 Objectives

The objective of this thesis is to explain employment and labor force growth by engaging non-traditional data sources. Instead of focusing on established determinants, like information about companies or political policy, this thesis investigates employment growth in relation to latent structure in a broader social environment. This structure reveals clearly recognizable trends in every day life, and is discussed in detail. These observations are then turned towards explaining employment growth and other related metrics.

1.2 Background

Measures of the economy and of economic growth have inspired much research. Within this, measures like total employment and job creation are covered regularly in great detail by the likes of the United States Bureau of Labor Statistics (BLS) and the Journal of Labor Economics. These issues also regularly make front page news, can dramatically impact major policy decisions, and move the stock market.

A fundamental indicator in this area of study is employment growth. Here, we investigate this, as well as five other labor force metrics. Employment growth, also called job creation, is an economy-wide figure, representing how many new paid positions were created in a given time period. We also look at labor force participation rates. These indicate the proportion of people over 16 who are working or actively looking for work. All employment or unemployment percentages, here and elsewhere, are percentages of the size of the labor force. The four other metrics considered are for subsets of economy-wide employment, and cover four main economic sectors, as discussed in Section 1.4.

To find determinants of job creation, researchers have typically analyzed firm-level characteristics [16, 21], investigated government policy[24, 25], or explored fluctuations within a specific industry [17].

Some researchers have attempted to comprehensively model employment growth, both from a micro- and a macroeconomic point of view [14, 15, 24]. This approach entails examining multiple major facets of social, demographic, and economic data, and exploring correlations with the various employment and labor force participation metrics. However, this research has generally not been re-evaluated since the 1980's, and, at the time, faced severe analytical restrictions. The study presented here will revisit this approach with more advanced computing techniques, as well as more recent, and much more comprehensive data.

Separate from the research on employment growth, there is a body of literature that focuses on describing complex latent structures found in social, economic, and demographic data. These works use statistical techniques such as principal component analysis (PCA) [20], k-means clustering [4, 8], or Ward's hierarchical clustering [8] to discover and interpret underlying patterns of variation. This work is entirely descriptive, and in two of the cases, is explicitly targeted towards selling business-to-business market research. Additionally, no connections are investigated between these results and other economic indicators.

This thesis will bridge the gap between these two fields of research. First, by applying principal component analysis to find and describe complex latent structures in United States Census data at the county-level. Second, by using the findings as independent variables in linear regression models to establish relationships with employment growth metrics.

This novel approach is strengthened by the greatly enhanced interpretability provided by PCA, as well as by the clear delineation of revealed groupings within the socio-economic and demographic variables. Together, these significantly clarify the results of linear regression analysis.

1.3 Data

All data in this study come from the American Community Survey (ACS), which is collated by the United States Census Bureau. The ACS is a nationwide survey that collects data on social, economic, housing, and demographic characteristics from 3.5 million households annually [6]. The ACS releases three types of statistical estimates each year: 1-, 3-, and 5-year estimates. All data used herein are 5-year estimates. These estimates are generated from 60 months of collected data, and are provided for every county. The Census recommends that these data be used when ”precision is more important than currency” [7].

All data used are at the county or county-equivalent level. There are 3,142 counties and their equivalents across the United States, of which 3,108 are in the lower 48 states. This work does not cover Alaska, Hawaii, or Puerto Rico, primarily because of their geographic isolation. The relatively small populations of Hawaii and Alaska, (40th and 48th by population respectively) also mean that their exclusion should not heavily bias the final results.

Of the 3,108 counties examined, 106 are ’county equivalents’ [10]. These are areas that are similar in nature to counties, but are given special designations for historical or statistical reasons [11], and fall into the following three categories. Louisiana is divided across 64 parishes[9], there are 41 ’independent cities’ across four states [11], and the District of Colombia is treated as its own county-equivalent [11].

Across the ACS database, there are over 1,000 unique subsets of 5-year estimates published annually for counties and county-equivalents [10]. Of these, the most commonly used are the four Data Profiles [5]. These Profiles provide summaries aimed at covering the most basic data across all topics [5]. Related, but more detailed data, is provided by 69 subdivisions called Subject Tables [5, 10]. All of the main data used in this paper come from these Subject Tables.

Variables that were subsets or cross-sections of the variables in Table 1.1 were excluded. Variables with values split out by race, ethnicity, or sex were aggregated.

Table 1.1: Subject Table Variable Names (Product Code)

Age and Sex (S0101)	Employment Status (Weeks Worked) (S2301)
Commuting Characteristics (S0801)	Work Status (S2303)
Households and Families (S1101)	Marital Status (S1201)
Industry (S2403)	Educational Attainment (S1501)
Field of Bachelor’s Degree for First Major (S1502)	Income (S1901)

Total employment growth is derived from Work Status data. Employment growth across sectors is calculated using the Industry data, and is divided into four sectors. Broadly speaking, these sectors are: raw materials extraction; manufacturing and construction; general services; and knowledge-based services [23]. These sectors are explored in more detail in the next section.

1.4 Employment Growth

Employment growth and related metrics are the main focus of this work. We look both at growth in actual employment, as well as growth in labor force participation. We also investigate employment growth within general economic sectors. There are four generally accepted sectors. These are referred to as the primary, secondary, tertiary, and quaternary sectors, and are described in detail by Zoltan Kenessey of the U.S. Federal Reserve [23]. These sectors divide all economic activity into categories of related pursuits, and are detailed below.

Table 1.2: Industries by Sector

Sector	Activities (SIC Code)
Primary	Agriculture, forestry, and fishing; Mining (1, 2, 7-14)
Secondary	Construction; Manufacturing (15-17, 20-39)
Tertiary	Transportation, electric, gas, and sanitary services; Wholesale trade; Retail trade (40-59)
Quaternary	Finance, insurance, and real estate services; Higher education; Public administration (60-67, 70, 72, 73, 75, 76, 78-89, 91-97)

In ACS data, sanitary services are grouped with administrative and support services and cannot be separated. Therefore the table above will be used with the exception of placing waste management services in the Quaternary sector.

1.4.1 General Employment Growth

Average employment growth at the county-level in the US is consistently near zero from 2009 to 2017. The same is true for growth in labor force participation; see Table below.

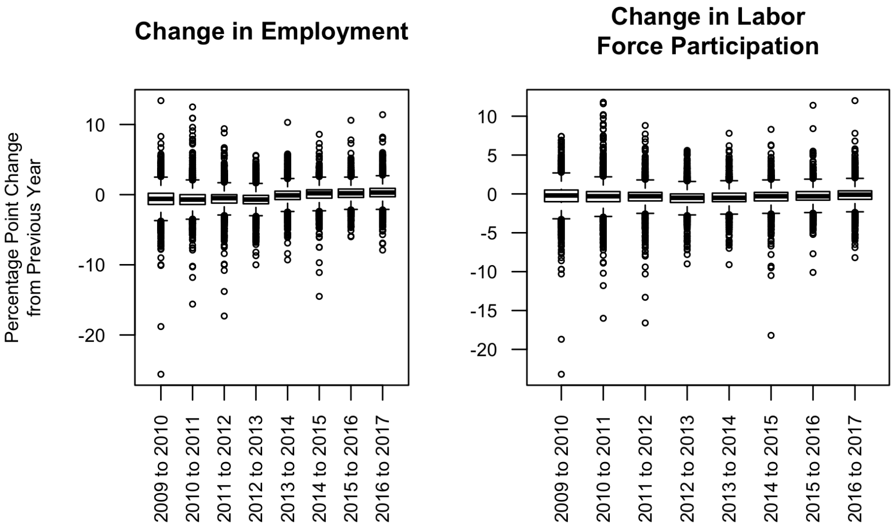
Table 1.3: Distribution of Average Percentage Point Changes in Employment and Labor Force Participation

	Average Percentage Point Change	
Year	**Employment**	**Labor Force Participation**
2009 to 2010	-0.599%	-0.285%
2010 to 2011	-0.676%	-0.346%
2011 to 2012	-0.579%	-0.335%
2012 to 2013	-0.676%	-0.514%
2013 to 2014	-0.110%	-0.441%
2014 to 2015	0.084%	-0.395%
2015 to 2016	0.162%	-0.280%
2016 to 2017	0.268%	-0.147%

No year had more than a one percentage point change from the previous year. In addition, labor force participation rates are all negative, as are the employment figures for five of the eight years examined. Values near zero are not unexpected, as the United States is a mature economy with low overall economic growth. These growth rates, however, are also likely due in part to the lingering effects of the 2007-2008 financial crisis. Note that change in employment is negative until 2014 before returning to positive growth.

Though centered roughly around zero, there is major variation around these means, with many counties experiencing more than a five percentage point gain or loss from one year to the next. In extreme cases, this change can be greater in magnitude than ten percentage points, see Figure 1.1.

Figure 1.1: Distribution of Changes in Employment and Labor Force Participation



These deviations can be more easily seen when displayed geographically, but it is also clear that there are not major geographic concentrations for either variable. Therefore, we must look deeper to find what is driving these growth rates.

Figure 1.2: Employment Growth (2016 to 2017)

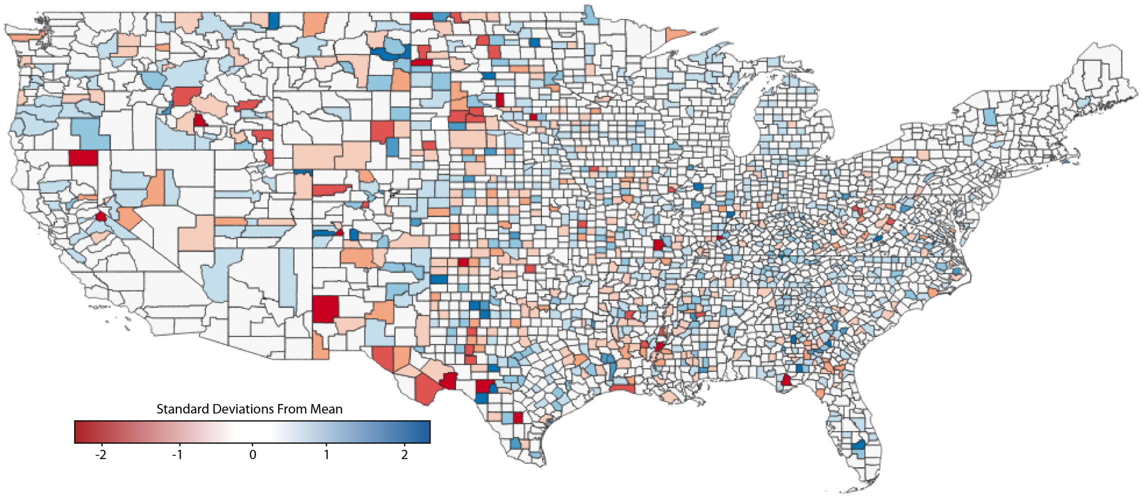
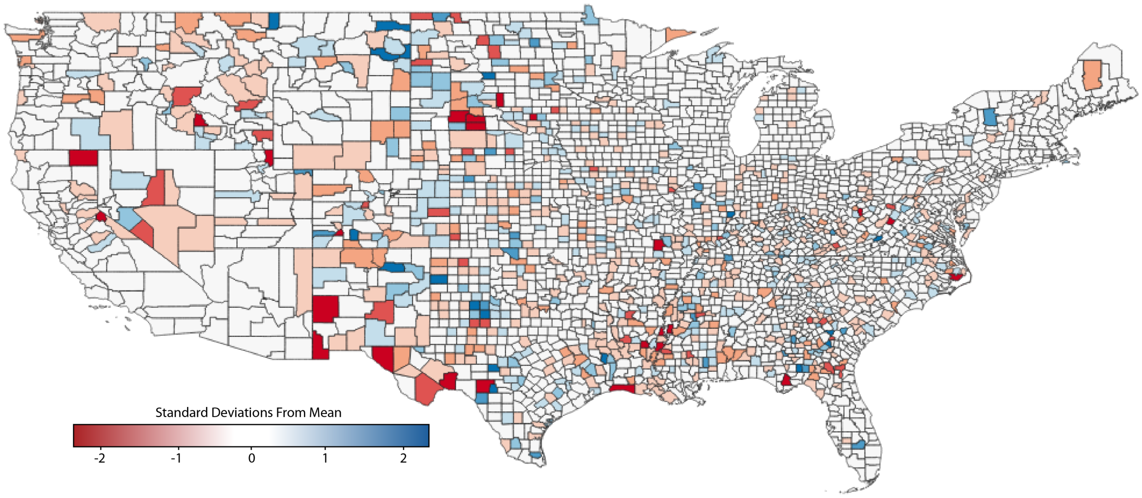


Figure 1.3: Labor Growth (2016 to 2017)



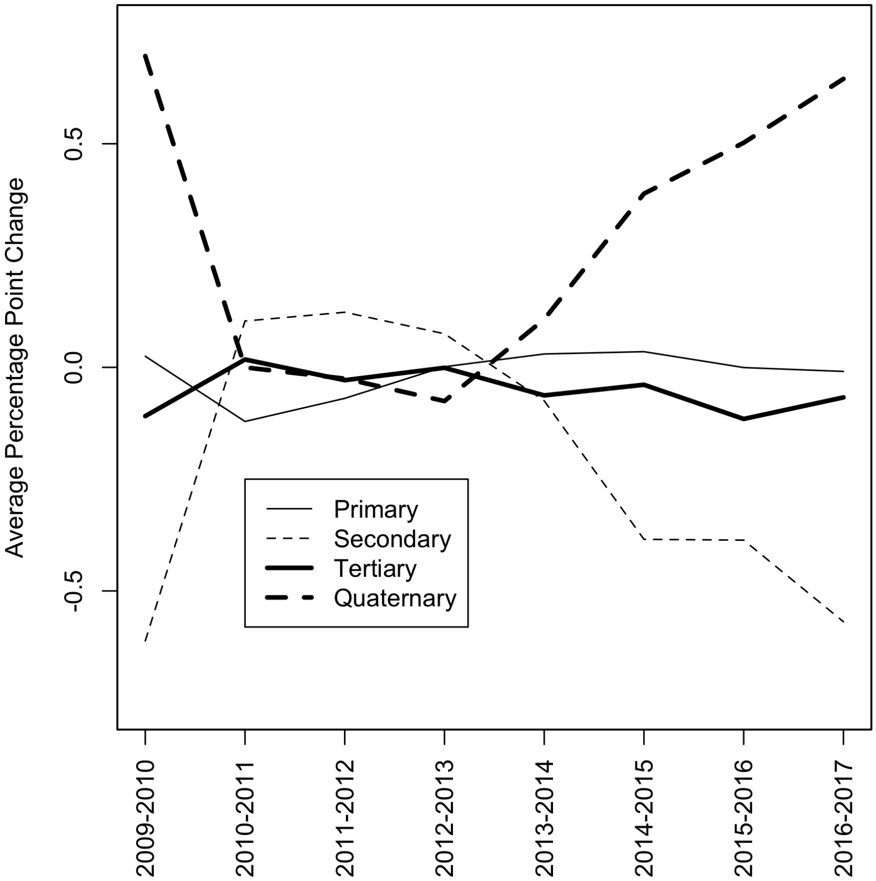
1.4.2 Sectoral Employment Growth

General employment measures are important, but there is value in examining the different types of activity within the labor force as well. To do this, industries were grouped into four sectors as specified above, and their growth rates investigated.

Table 1.4: Distribution of Average Percentage Point Changes in Employment By Sector

	Average Percentage Point Change			
	Primary	Secondary	Tertiary	Quaternary
2009 to 2010	0.025%	-0.612%	-0.109%	0.696%
2010 to 2011	-0.121%	0.104%	0.018%	0.000%
2011 to 2012	-0.069%	0.123%	-0.029%	-0.025%
2012 to 2013	0.001%	0.075%	-0.001%	-0.075%
2013 to 2014	0.030%	-0.075%	-0.063%	0.108%
2014 to 2015	0.035%	-0.385%	-0.039%	0.388%
2015 to 2016	0.000%	-0.386%	-0.115%	0.502%
2016 to 2017	-0.009%	-0.569%	-0.067%	0.645%

Figure 1.4: Average Sector Growth Across Years

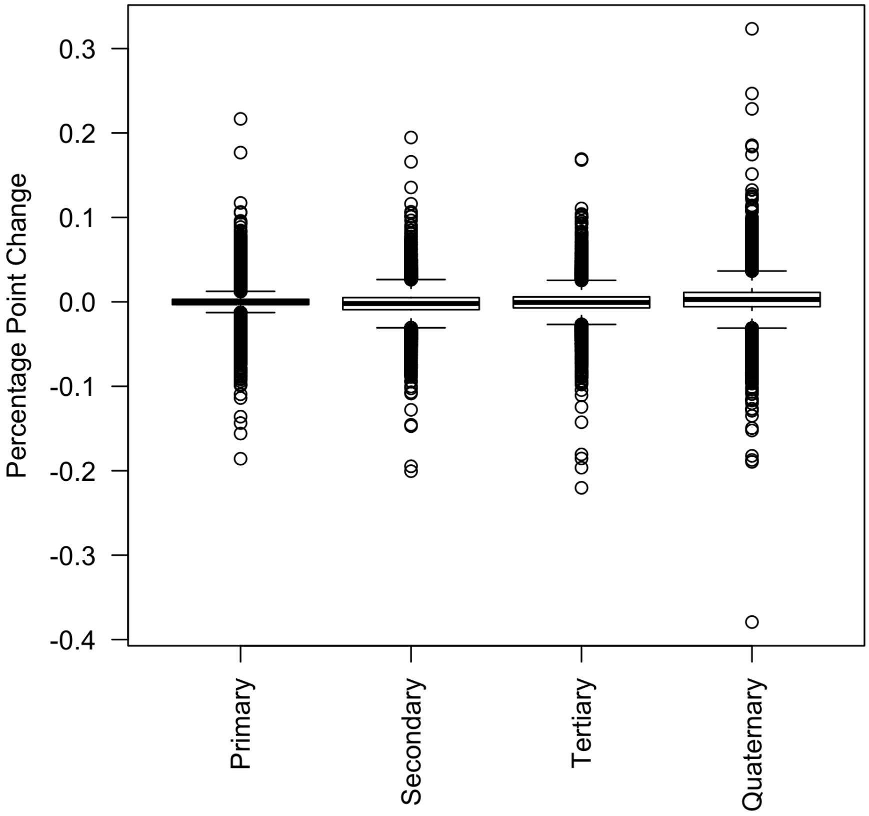


Again, these growth rates are close to zero, with no average exceeding one percentage point in magnitude. This is expected based on the general analysis above, though more subtlety can now be observed.

The primary and tertiary sectors hew closely to this zero growth trend throughout. This is plausible, as the primary sector consists of resource extraction, and the tertiary covers basic services and all types of retail. These are in constant and inelastic demand, and therefore, would not be expected to react strongly to changes in the overall economic environment.

Secondary sector economic activities increase markedly at the beginning of the time frame, as shown in Figure 1.4, but from a strong negative rate of change. The secondary sector covers construction and manufacturing. The former was intimately involved in the 2008 financial crisis [18], so it is perhaps not surprising that the sector was recovering from a period of negative growth. Growth in the quaternary sector dropped off precipitously after the end of the financial crisis, but then began to grow again after a couple of years. Greater fluctuations in these sectors is expected. The manufacturing and construction industries make more durable goods, and so are subject to greater fluctuations in demand. The quaternary sector is made up of information services. These can be extremely valuable, but are more likely to enhance a business rather than be fundamental to its operation. Therefore this sector too would be expected to be more susceptible to economic trends. The quaternary sector also employs the majority of workers in the United States, so general fluctuations in the economy will likely affect this sector in a similar manner.

Figure 1.5: Annual Growth Rates Across Sectors



Interestingly, there is very little difference in growth rate variation across sectors, though the national year-on-year averages shown above do fluctuate.

Chapter 2

Methods

In pursuit of predicting the employment growth metrics in Section 1.4, the following methods are employed. First the data was preprocessed to ensure validity and standardize the variables. Then principal component analysis and linear regression were used on this cleaned data. All work was done in *R* (Version 3.3.3), using the *base*, *dplyr*, and *car* packages [26].

2.1 Data Preprocessing

2.1.1 Census Data: Margins of Error and Coefficients of Variance

Data from the American Community Survey comes in the form of a point estimate and a margin of error (MOE). The MOE is a measure of possible estimate variation around the 'true' population value, and is calculated using the variance estimated for the original estimate [19].

$$\text{Margin Of Error} = 1.645 * \text{Standard Error} \tag{2.1}$$

The number 1.645 is the Z-score for the 90% confidence level that is standard for the U.S. Census [19].

To determine if a data point is significantly different from zero, the MOE and the estimate can be used to calculate a coefficient of variation (CV) [1].

$$\text{Coefficient of Variation} = \frac{\text{Margin of Error}/1.645}{\text{Estimate}} \tag{2.2}$$

Here any coefficient over 40% is rejected as not statistically distinguishable from zero, using guidance from the Environmental Systems Research Institute (Esri)[1].

2.1.2 Data Transformation

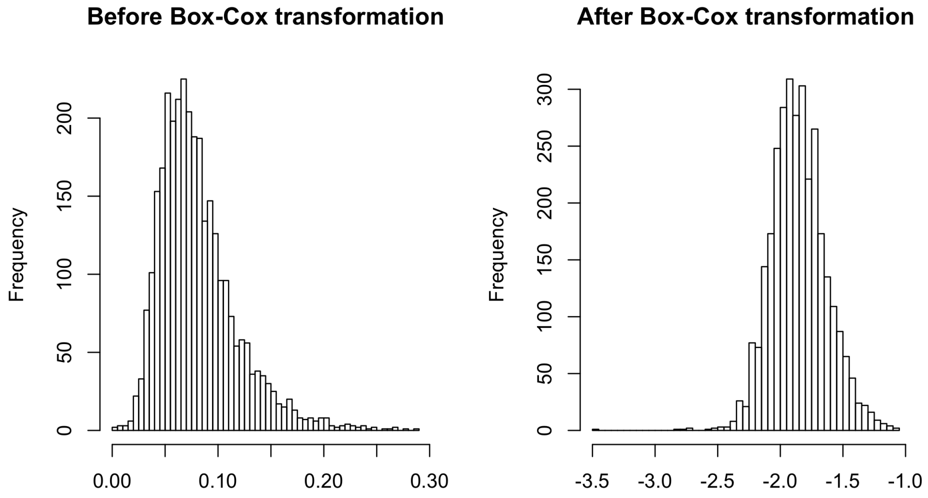
All data used in predicting employment growth metrics are transformed and standardized before further analysis. The data are transformed using the Box-Cox power transformation. The general form of this transformation is given by the function below. All power transformations require a value for the parameter λ. This is chosen to be the value that maximizes the function $\ell(\lambda)$, given by Equation 2.4.

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases} \tag{2.3}$$

$$\ell(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} \left(x_j^{(\lambda)} - \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)} \right)^2 \right] + (\lambda - 1) \sum_{j=1}^{n} \ln(x_j) \tag{2.4}$$

This function is continuous in λ for $x > 0$ [22], so optimization is possible. The Box-Cox transformation is designed to generally improve approximation to normality. In these analyses however, this is almost never achieved. Instead, the value of a power transformation is to scale the data towards a less skewed distribution, even if normality itself is not achieved. This improvement can be seen in Figure 2.1.

Figure 2.1: Effects of a Box-Cox Power Transformation on Example Income Data



After transformation, each variable is standardized by subtracting the mean μ and dividing by the standard deviation σ. This gives a column of data that is centered around zero, and has a standard deviation of one. Rescaling coerces each variable into a comparable range and variance to aid in clear analysis of potential effects.

$$x_{standardized} = \frac{x - \mu}{\sigma} \tag{2.5}$$

2.2 Principal Component Analysis

Principal component analysis aims to explain the covariance structure of a data set through linear combinations of the original variables. It is frequently a useful initial step before conducting other analyses [22]. The linear combinations generated are called the principal components of the data. To construct these, let $\mathbf{\Sigma}$ be the covariance matrix of a set of n observations across p variables, denoted $\boldsymbol{X}$. Let the eigenvalue-eigenvector pairs of $\mathbf{\Sigma}$ be denoted $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$, where the eigenpairs are ordered by decreasing eigenvalues.

The i^{th} principal component is:

$$Y_i = \boldsymbol{e}_i^T \boldsymbol{X}, \quad i = 1, 2, ..., p \tag{2.6}$$

With:

$$Var(Y_i) = \boldsymbol{e}_i^T \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i \quad i = 1, 2, ..., p \tag{2.7}$$

$$Cov(Y_i, Y_k) = \boldsymbol{e}_i^T \boldsymbol{\Sigma} \boldsymbol{e}_k = 0, \quad i \neq k \tag{2.8}$$

A given Y_i is the linear combination of the original variables needed to construct the i^{th} principal component. The coefficients for this linear combination are called the 'loadings' or 'rotations' for that component. The proportion of variance in $\boldsymbol{\Sigma}$ explained by the i^{th} component is given by (2.9).

$$\text{Proportion of Variance Explained}\,(Y_i) = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p} \tag{2.9}$$

This construction generates principal components that are both uncorrelated with each other, and that are ordered in terms of their ability to explain variation within the original data set $\boldsymbol{X}$ [22]. This transforms a potentially complicated correlation or covariance structure into a more simplified set of new variables.

This has three particular benefits. First, if much variability is explained by only a few principal components, then dimension reduction can be achieved by replacing the original data with these high value components in future analyses. Second, if the linear combinations found are clearly interpretable, then these descriptions can give meaningful insight into latent structures not clearly visible in the original data. Finally, working solely, or predominantly, with variables that have no correlation can greatly ease statistical analysis.

The first two of these benefits can be seen when examining 2016 income data for counties in the United States.

Table 2.1: Annual Income Brackets (2016)

Less than $10,000	$10,000 to $14,999
$15,000 to $24,999	$25,000 to $34,999
$35,000 to $49,999	$50,000 to $74,999
$75,000 to $99,999	$100,000 to $149,999
$150,000 to $199,999	$200,000 or more

Given these ten income brackets, principal component analysis finds the linear combinations shown in Table 2.2. Note, only coefficients over 0.1 are shown here, as values lower than this contribute little to the overall interpretation or component construction. Most latter analyses have fewer variables than used here, so only coefficients over 0.2 will be generally reported.

Table 2.2: Principal Component Analysis of Income Brackets (2016)

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**	**Comp.6**	**Comp.7**	**Comp.8**	**Comp.9**	**Comp.10**
Less than $10,000	-0.328	0.322	-0.101	0.263		0.611	-0.251	-0.317	-0.1	-0.396
$10,000 to $14,999	-0.353	0.165	-0.156	0.132		-0.573	-0.544	0.29	0.176	-0.259
$15,000 to $24,999	-0.37		-0.106			-0.248	0.756	-0.108	0.176	-0.412
$25,000 to $34,999	-0.306	-0.196		-0.841	0.238	0.16	-0.141			-0.223
$35,000 to $49,999	-0.176	-0.527	0.732	0.314						-0.21
$50,000 to $74,999	0.144	-0.608	-0.516		-0.514				0.1	-0.227
$75,000 to $99,999	0.324	-0.223	-0.158	0.163	0.711	-0.126	-0.135	-0.412	0.144	-0.252
$100,000 to $149,999	0.384							0.522	-0.519	-0.536
$150,000 to $199,999	0.359	0.223	0.16			0.249		0.258	0.779	-0.217
$200,000 or more	0.323	0.274	0.306	-0.267	-0.387	-0.357	-0.102	-0.536	-0.135	-0.251

The linear combination found for the first component has negative loading values for income brackets under $50,000 annually, and positive values for brackets over $50,000. The interpretation of this is an axis where values in the negative direction indicate a county with more households earning lower incomes, and where more positive values are found for a county with higher incomes. This understanding reveals broad generalizations about the underlying structure of patterns in income data that are not necessarily evident when working with the original variables. In addition, Table 2.3 below shows that movement along this axis, or component, accounts for 56.4% of all variation found in the raw data covariance matrix $\mathbf{\Sigma}$. This means that this one new variable could replace the ten input variables while still keeping a slight majority of the original variability. Using the first two components we account for 56.4% and 15.9% of variation respectively, explaining a total of nearly three quarters of the original variability while reducing the dimensionality of the data set by four-fifths.

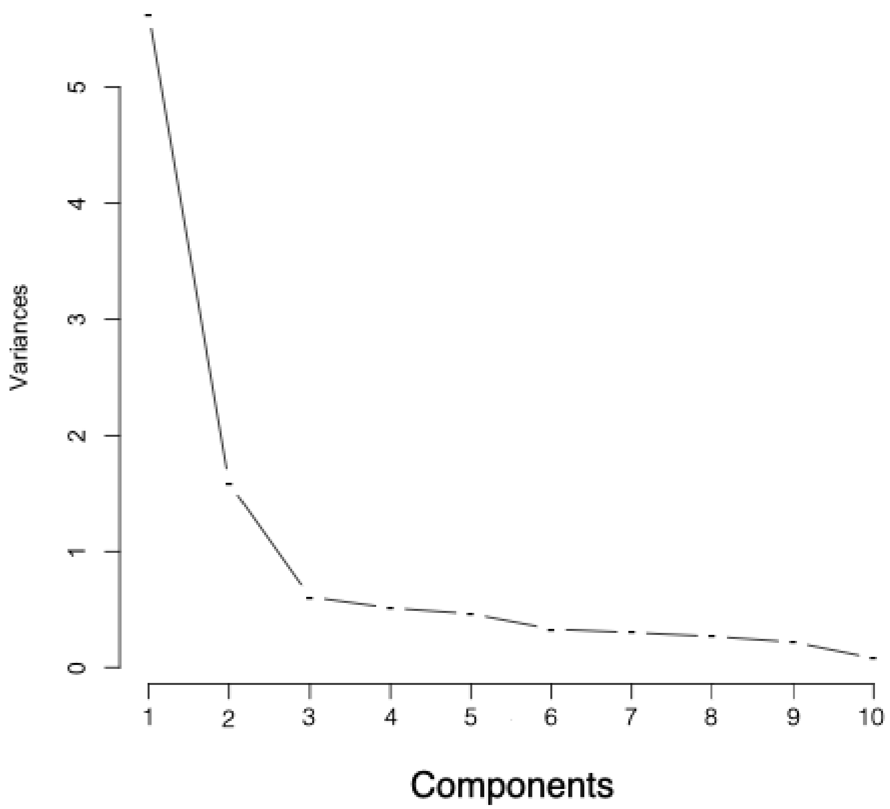
Table 2.3: Income Principal Component Variances (2016)

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**
Standard deviation	2.3735	1.2598	0.7794	0.7156	0.6800
Proportion of Variance	0.564	0.159	0.061	0.051	0.046
Cumulative Proportion	0.564	0.722	0.783	0.834	0.881
	Comp.6	**Comp.7**	**Comp.8**	**Comp.9**	**Comp.10**
Standard deviation	0.5751	0.5480	0.5166	0.4668	0.2804
Proportion of Variance	0.033	0.030	0.027	0.022	0.008
Cumulative Proportion	0.914	0.944	0.970	0.992	1

Using all ten components will yield a data set containing all of the variability in the original ten income brackets, but the latter seven components each account only for a small and decreasing proportion of this variation.

Choosing the ultimate number of components to keep, however, is not a rigorously defined process. A 'scree plot' can aid in this, with the goal of identifying a set of high contributing components; see Figure 2.2.

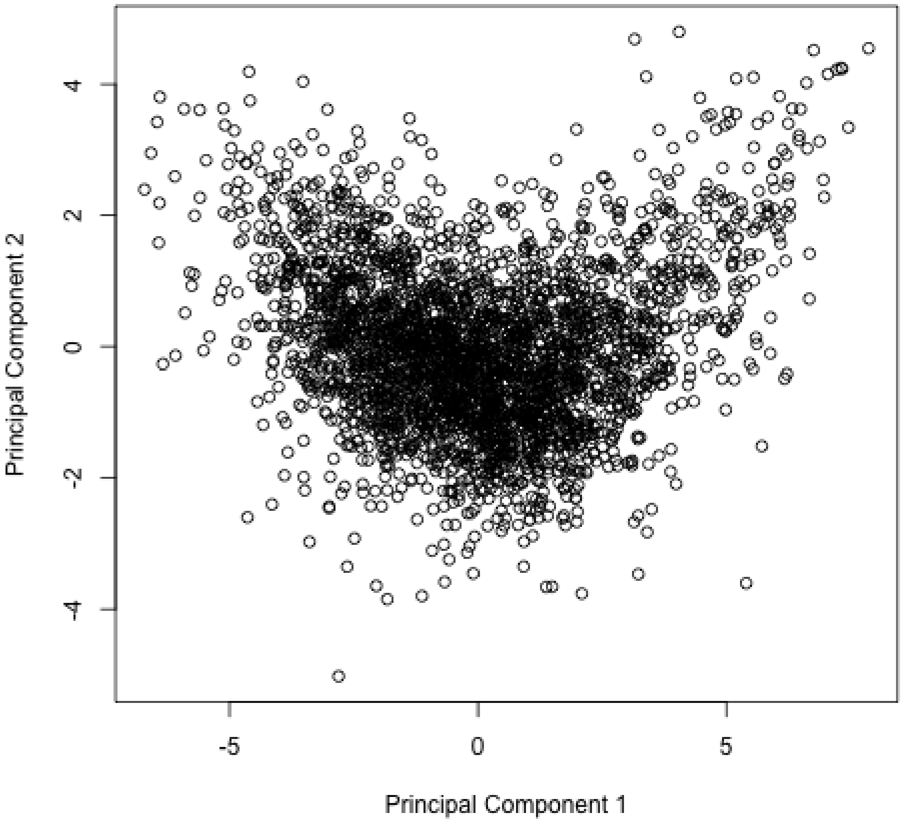
Figure 2.2: Example Scree Plot of Component Variances



Here there is a clear change in variance explained starting at the third component. This is often referred to as the 'elbow' of the plot, and is recommended as an indicator of the maximum number of useful components [22]. Interpretational understanding is also a helpful potential indicator.

Principal component analysis does not require a multivariate normal assumption, instead relying solely on the covariance or correlation matrix of the relevant variable [22]. Though multivariate normally distributed data are not required, most component quality checks do rely on this property. The major exception is the examination of component versus component plots. Plotting each pairwise combination of components gives an opportunity to identify and investigate potential outliers or faulty data points. The plot below shows clear structure, due to the non-normality of the data, but also that there are no wildly divergent outliers.

Figure 2.3: Example Principal Component Comparison For Income Components 1 and 2



Performing common quality checks such as χ^2 ellipses, quantile-quantile plots, or T^2 threshold checks will not return useful information for these components due to this structure.

Finally, in calculating the components, I used only the covariance matrix. This is because the data is transformed and standardized before analysis, and so the coercion to comparability gained by using the correlation matrix is not necessary.

2.3 Bootstrapping

All principal component analyses are checked for robustness and sensitivity via the bootstrapping technique. Boostrapping involves creating a 'new' data set by sampling with replacement from the observations in the original dataset. By testing many different subsets of the data, a clearer picture of the underlying distribution of the data can emerge. Here, this is used to understand the sensitivity of the principal component construction to changes in the data. The average of the component loadings, plus 90% confidence intervals around these values, are calculated using 5,000 resamples. These are then plotted across the study period from 2009 to 2016. This shows if the values are stable across time as well as stable with respect to resampling. Though the confidence bounds are printed on all of these plots, they are not always visible. This is because the confidence interval around the point estimator is much smaller than differences between different levels plotted. This true even though the plots are either percentages or principal component loadings, with the latter rarely spanning more than 2 'units'.

2.4 Linear Regression

After principal component analysis has been performed and investigated, the components are used as independent variables in linear regression analysis. Linear regression uses these variables to predict a response, or dependent variable. This dependent variable is denoted by $\boldsymbol{y}$, and the independent variables are represented by the matrix $\boldsymbol{X}$. This is an $N \times N$, matrix where N is the number of independent variables. A column vector of ones is included in the first column to allow for an intercept term to be estimated as part of the linear function.

The model for linear regression then, is shown in (2.10).

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.10}$$

The vector $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad ... \quad \beta_N]$ is a set of coefficients that specify the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$, and includes an intercept term β_0. The $\boldsymbol{\epsilon}$ is the difference between the observed values of $\boldsymbol{y}$, and those estimated by the hypothetical linear model. Any actual implementation of linear regression will not find this true $\boldsymbol{\epsilon}$. Instead this vector is approximated by the estimated errors $\hat{\boldsymbol{\epsilon}}$, also known as the residuals.

The true linear coefficients $\boldsymbol{\beta}$ are also not observable, and so are estimated using the ordinary least squares (OLS) method. These estimations are denoted $\hat{\boldsymbol{\beta}}$.

OLS seeks to minimize the sum of the squares of the residuals, or estimated errors, between the known and estimated values [22]. This is done by solving the following equation, where $\hat{\boldsymbol{\beta}}$ is the vector of estimates for the hidden true $\boldsymbol{\beta}$ coefficients.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.11}$$

An estimate of the dependent variable is then generated for each observation by multiplying $\hat{\boldsymbol{\beta}}$ by the independent variables $\boldsymbol{X}$. This estimate is also known as a predicted value, or a fitted value, and is represented as $\hat{\boldsymbol{y}}$.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{2.12}$$

With $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{\epsilon}}$ can be calculated.

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{2.13}$$

These values can now be used in concert to estimate the relationships between $\boldsymbol{X}$ and $\boldsymbol{y}$, however, care must be taken to investigate each model to confirm its' veracity. First, linear regression is susceptible to high degrees of correlation between predictor variables (called multicollinearity). To avoid this, all variables selected should be at least not perfectly correlated with each other. This is equivalent to requiring that the matrix $\boldsymbol{X}$ have full rank.

Each linear model must also be checked for validity after estimation. One possible problem is correlation between the residuals and each of the variables in $\boldsymbol{X}$. This can be checked by plotting the residuals against each independent variable. If clear patterns are observed with respect to values for the independent variable, this may suggest more terms are needed in the model. A second issue is non-constant variance or a non-normal distribution of the residuals. Presence of either of these may indicate influential outliers or possibly data with a particularly abnormal distribution. In addition to these two properties, it is imperative that the residuals are not correlated with each other [22].

Selecting a set of predictor variables for a model can also be a challenge. Here, forward-stepwise selection is employed. This technique involves starting with no independent variables, and then testing all possible single variable models. The model with the smallest residual sum of squares (RSS) is selected.

$$RSS = \sum_{j=1}^{n} \hat{\epsilon_j}^2 \tag{2.14}$$

Then all possible models of one additional variable are tested, and again, the model with the smallest RSS is selected. This process can then be repeated until the error for the next best model is above a certain threshold. However in these analyses, there are many possible independent variables, so no more than ten to fifteen possible models are checked.

From these, the model with the most explanatory power is chosen. To do this, cross-validation and the adjusted-R^2 statistic are used. Adjusted-R^2 is a variation on R^2, which is a measurement of the proportion of total variation in $\boldsymbol{X}$ explained by a given model. R^2 will increase whenever an additional variable is included however, regardless of the true explanatory value of that variable. Adjusted-R^2 attempts to control for this by accounting for the number of predictor variables used.

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \hat{\epsilon_j}^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} \tag{2.15}$$

$$Adjusted - R^2 = 1 - \frac{(1 - R^2)(n-1)}{n - r - 1} \tag{2.16}$$

Here $\bar{y}$ is the expected value of y, n is the number of observations, and r is the number of independent variables in $\boldsymbol{X}$.

Cross-validation is a technique that uses part of a dataset to test prediction accuracy. To do this, a subset of the data are set aside for testing, while the rest of the data are used to parameterize the model in question. Here, this means using 80% of the full data set to train each model, and then calculating the mean squared prediction error based on the remaining 20%. This 20% is usually between 150 and 600 observations, depending on the quality of a given data set.

$$MSE_{prediction} = \frac{1}{n} \sum_{a}^{b} (Y_i - \hat{Y_i})^2 \tag{2.17}$$

To keep the same level of statistical confidence across all analyses, a p-value of 0.1 is used to determine the significance of regression coefficients. This is in line with the 90% confidence intervals estimated by the American Community Survey.

Chapter 3

Descriptive Results

In this section, principal component analysis is used to uncover structures in the data described above. The variables in each data set are described and examined, before being deconstructed into their principal components. This deconstruction dramatically strengthens interpretation, while reducing the variable dimensionality. Also shown, is clear stratification of related trends within these variables. An in depth examination is conducted for the 2016 data, including geographic visualization. This is then put in the context of the entire study period (2009 to 2016) at the end of each section.

For datasets where many groupings are present, preliminary principal component analysis is used to explore the possibility of dimension reduction. Often this results in fewer variables being used in the main analysis than are originally collected by the Census Department. If this is the case, the original analysis is preserved in the Preliminary Analysis Appendix. Generally, only principal component loadings over 0.2 are reported for clarity.

3.1 Industry Sectors

In addition to looking at growth rates for the economic sectors, absolute figures can help shed light on major trends in the economy as well. Figure 3.1 shows the distribution of employment for each sector. In this plot, the box shows the interquartile range, with the median value in the middle. The tails show 1.5 times the interquartile range, and give an indication of the general distribution of the data away from the center. Any outliers that lie outside of this range are shown as individual points.

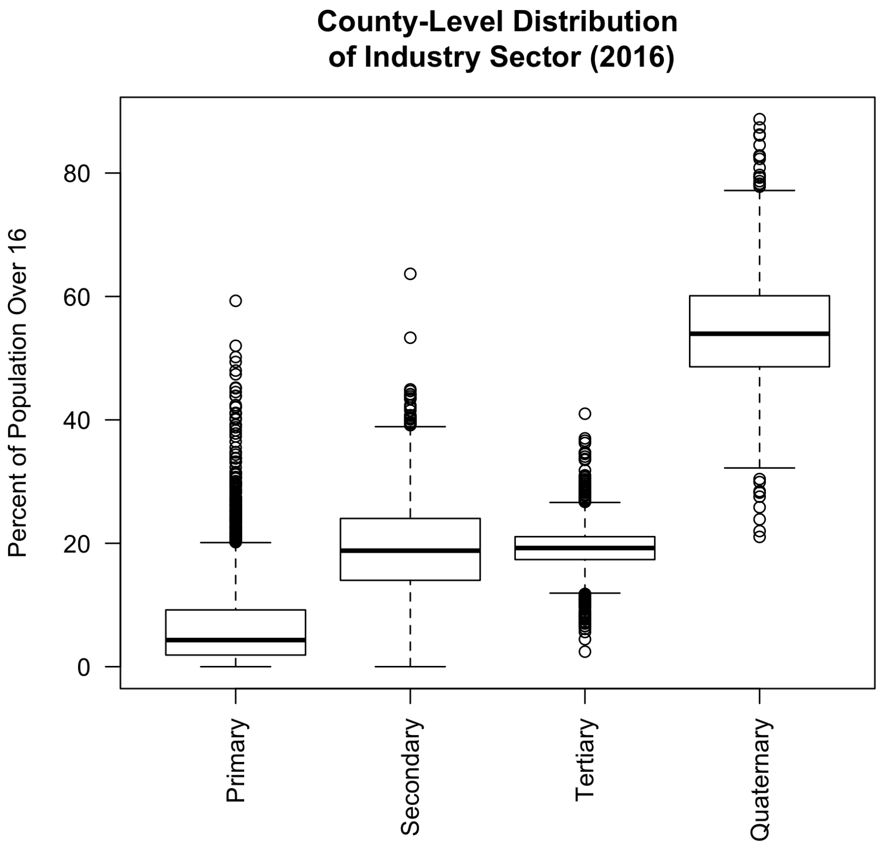
The first, or primary, sector covers resource extraction and agriculture. These industries underpin all other economic activity, but make up a small proportion of overall employment in the United States. Primary sector activities employ under 10% of workers in the average county, and is also an almost zero-growth set of industries. The secondary sector covers manufacturing and construction. This sector employs an average of around 20% of all workers, and is more susceptible to changes in economic outlook.

Tertiary activity is composed of basic services. These include transportation services, power generation, wholesale trade, and retail trade. On average a similar proportion of workers are employed in tertiary activities, but the variance is much lower. This is because basic services are essential to an advanced, modern, economy, and so are more uniformly present in each county. This unchanging, or inelastic, demand also contributes to the low-growth nature of the sector.

The final, or quaternary, sector, is made up of information- and high-technology services. These cover industries such as finance, insurance, higher education, and public administration, and dominate employment in the United States. The average county has more than half of its' workers employed in these industries, and no county has fewer than 20%.

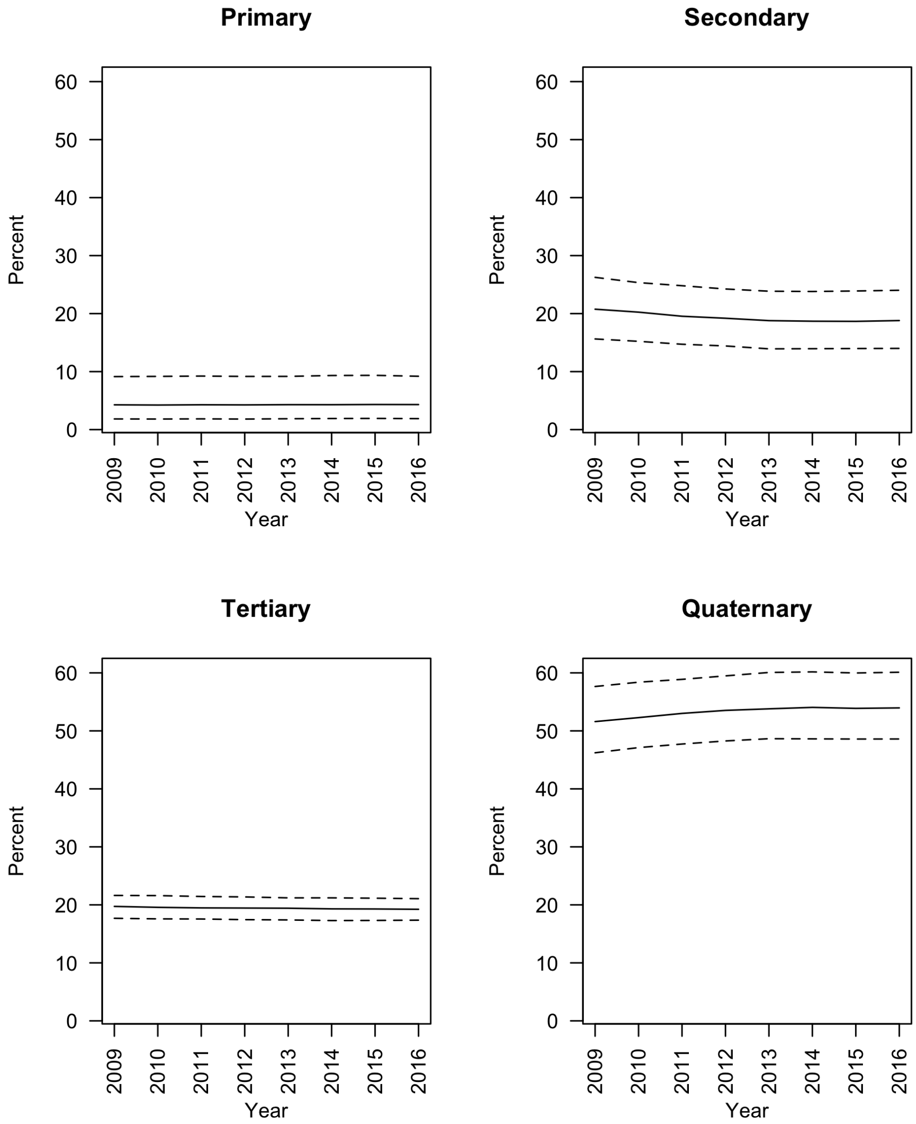
Information- and technology-based services are also more susceptible to market fluctuations. Employment in the sector dropped to zero in the years after the 2007-2008 financial crisis, but from 2013 to 2017, it has been the only portion of the economy to see broad and sustained growth.

Figure 3.1: Distribution of Sectoral Employment (2016)



Generally, trends across the sectors are mostly stable. A drop in secondary sector employment can be seen in the aftermath of the 2007-2008 financial crisis, as well as an uptick in employment in quaternary industries. However, these fluctuations are both well within their respective inter-quartile ranges. The relative dominance of quaternary activities can be seen here as well.

Figure 3.2: Inter-Quartile Ranges of Sectoral Employment



3.1.1 Principal Component Analysis

The strongest trend found across sector-level employment data is that counties tend toward quaternary sector jobs, versus all other sectors. The next strongest trend sets primary sector employment against secondary and tertiary industries. Finally, counties are distinguished between being more oriented towards tertiary versus secondary employment. Together, these three axes explain 98% of the variation in the original 4-variable data set.

Table 3.1: Sectoral Employment: Principal Component Loadings (2016)

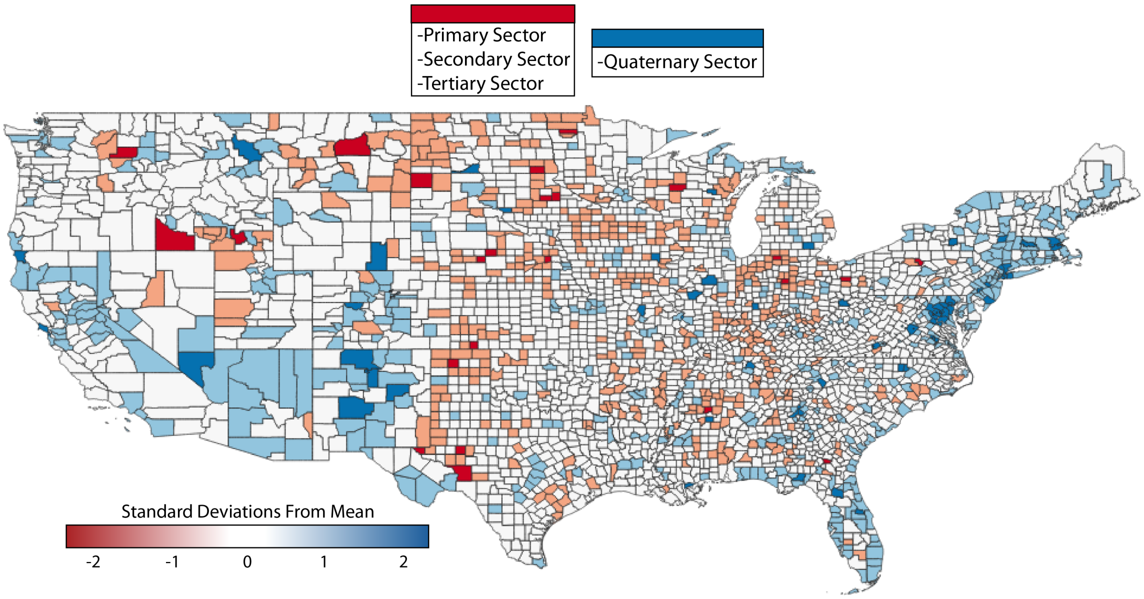
	Comp.1	**Comp.2**	**Comp.3**
Primary Sector	-0.464	0.702	
Secondary Sector	-0.427	-0.621	-0.444
Tertiary Sector	-0.174	-0.348	0.895
Quaternary Sector	0.757		

Table 3.2: Sectoral Employment: Principal Component Variances (2016)

	Comp.1	**Comp.2**	**Comp.3**
Standard deviation	1.2985	1.1132	0.9964
Proportion of Variance	0.422	0.31	0.248
Cumulative Proportion	0.422	0.732	0.98

The clearest concentration of quaternary employment is found in the Washington D.C. area, as well as along the metropolitan corridor to New York City. Industries in the quaternary sector are the main employers in the United States, so any county marked in blue has a high proportion of people employed in these areas, even relative to the country baseline.

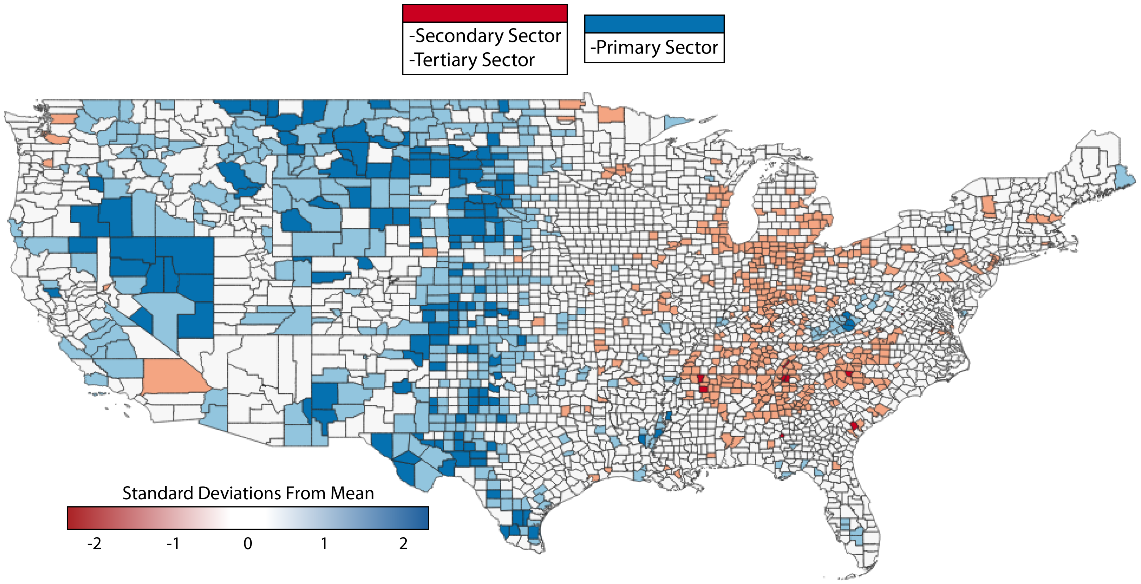
Figure 3.3: Sectoral Employment: Component 1 (2016)



The second sectoral component is more clearly geographically stratified than the first. The clear tendency towards primary sector jobs in the Western half of the country is indicative both of the major industries in the area, as well as of the generally lower population in rural Western counties. This is because counties with fewer people can be disproportionately employed by a single industry or firm.

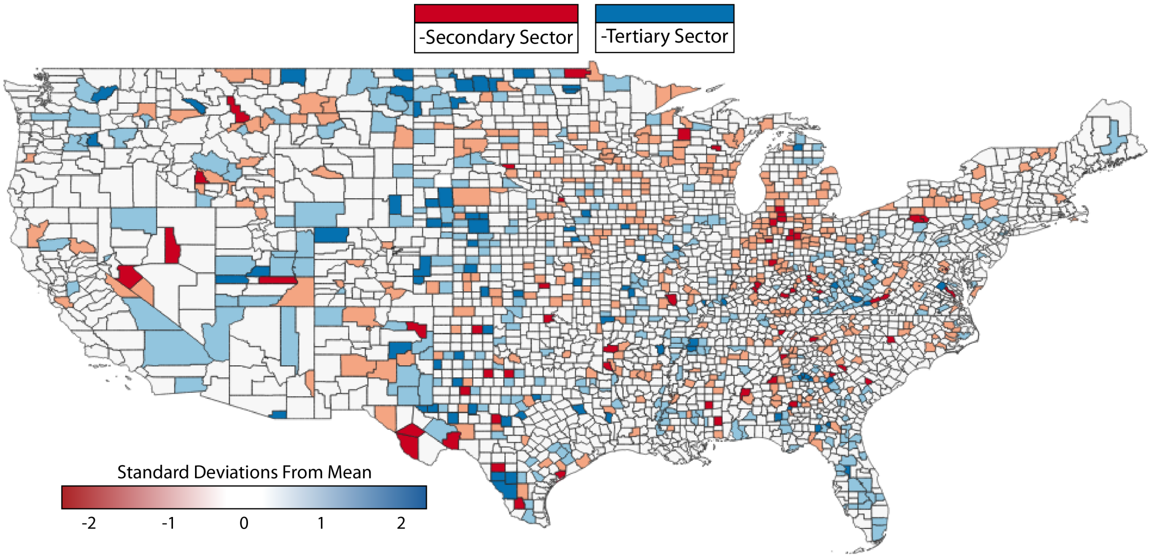
In the Eastern half of the country, population density is generally higher. This causes sectoral employment to be more evenly spread, but it also means that the pockets of secondary and tertiary employment observed represent on average a more dense concentration of people working in a certain set of industries.

Figure 3.4: Sectoral Employment: Component 2 (2016)



The third industry sector component differentiates between secondary and tertiary economic activities, and appears not to be geographically concentrated.

Figure 3.5: Sectoral Employment: Component 3 (2016)



The principal components found across sector employment data are very stable with respect to time and resampling, with no major reversals during the study period.

Figure 3.6: Sectoral Employment - Component 1: 2009 to 2016

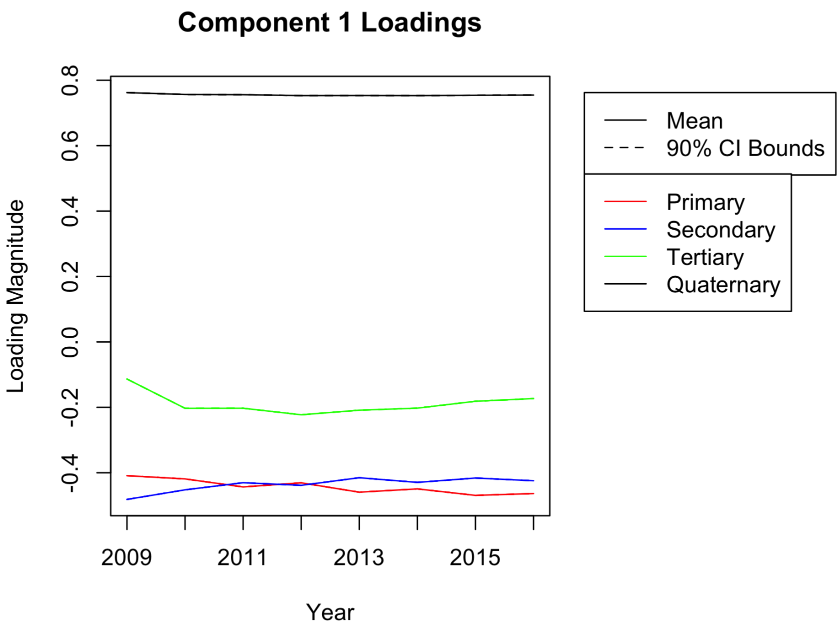


Figure 3.7: Sectoral Employment - Component 2: 2009 to 2016

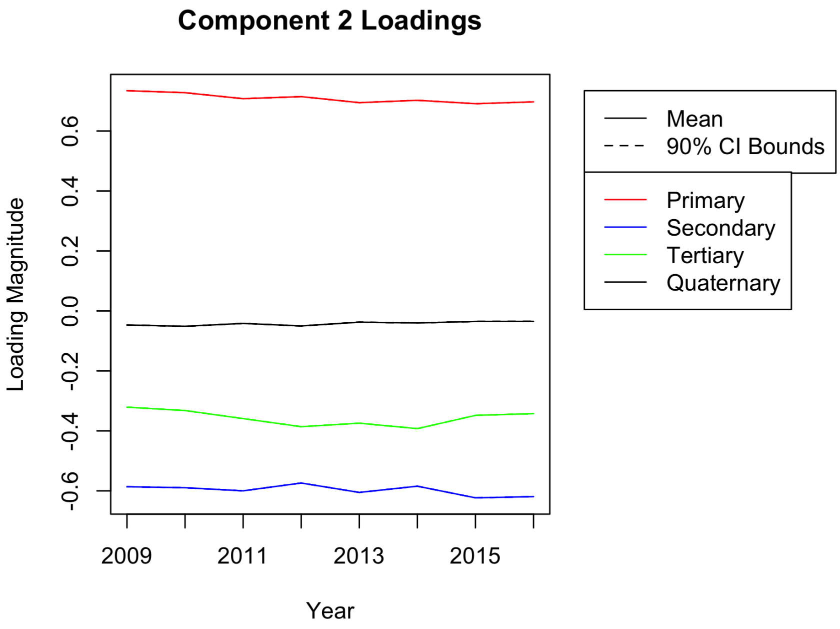


Figure 3.8: Sectoral Employment - Component 3: 2009 to 2016

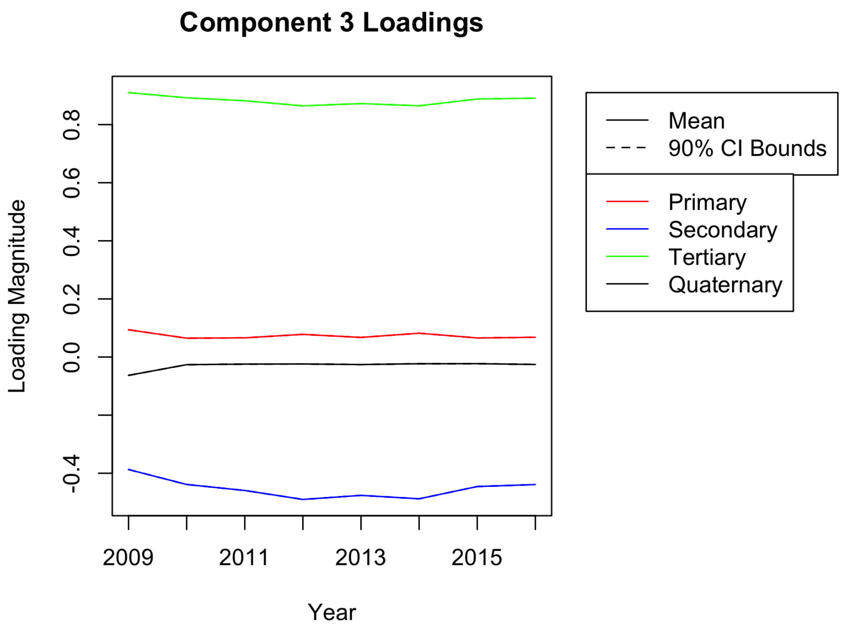
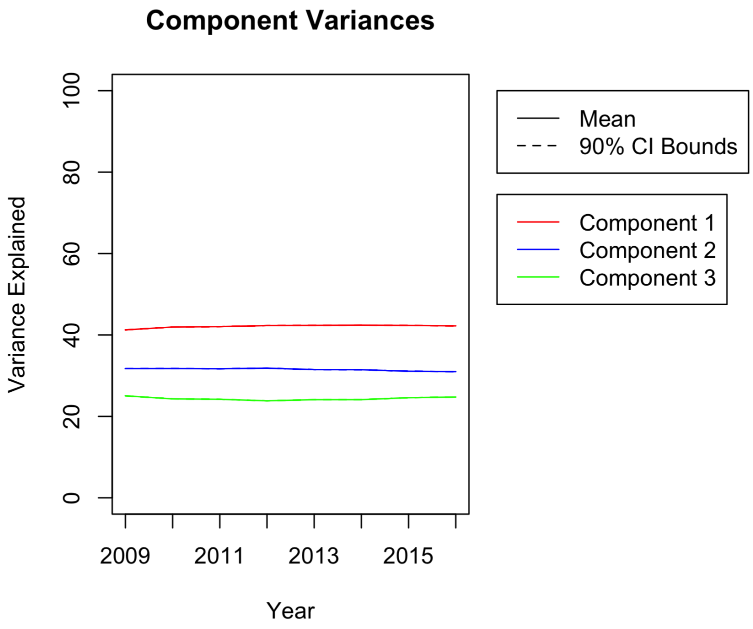


Figure 3.9: Sectoral Employment - Proportion of Variance Explained: 2009 to 2016



3.2 Income

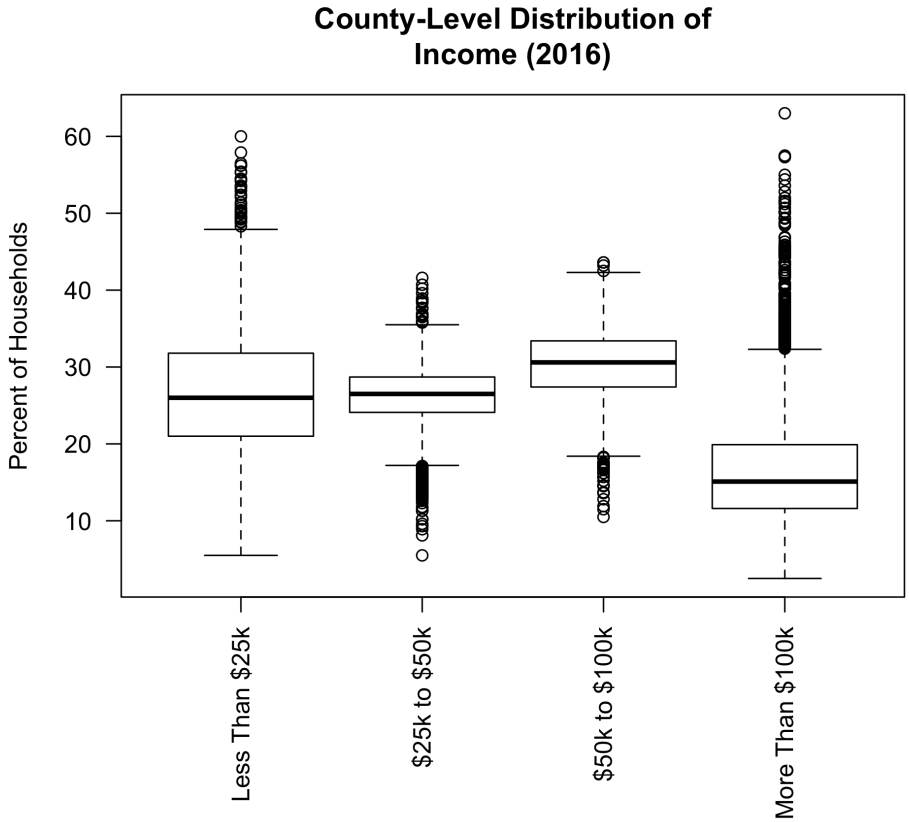
The Census collects household annual income data across 10 brackets. Exploratory principal component analysis contained in the Appendices shows that grouping these into four brackets maintains and strengthens the underlying patterns found across all 10. These four condensed income brackets are shown below.

Table 3.3: Income Brackets

Less Than $25k	$25k to $50k
$50k to $100k	More Than $100k

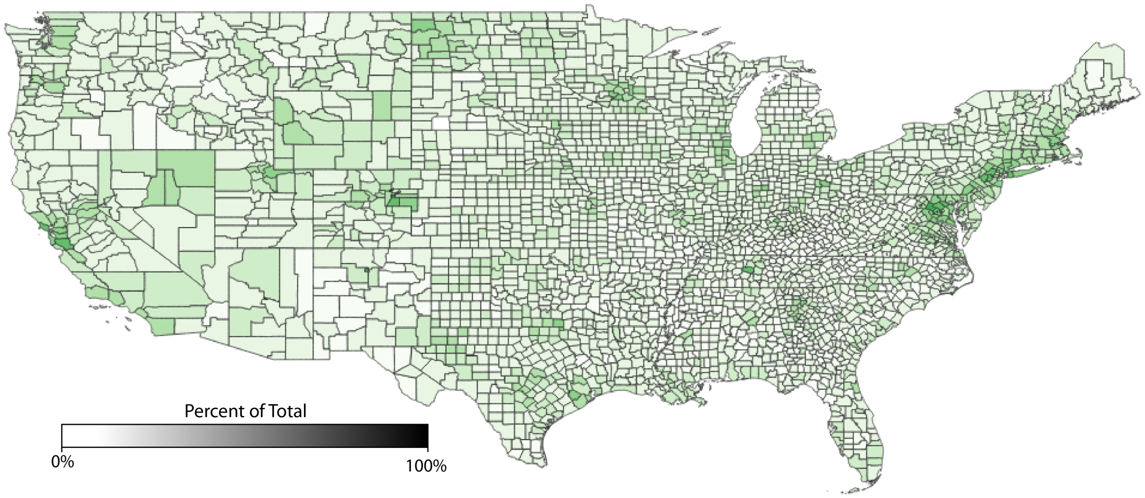
The lower three brackets do not much differ between themselves, though the percentage of people earning $100,000 or more per year is clearly less common on average.

Figure 3.10: Income Distribution (2016)



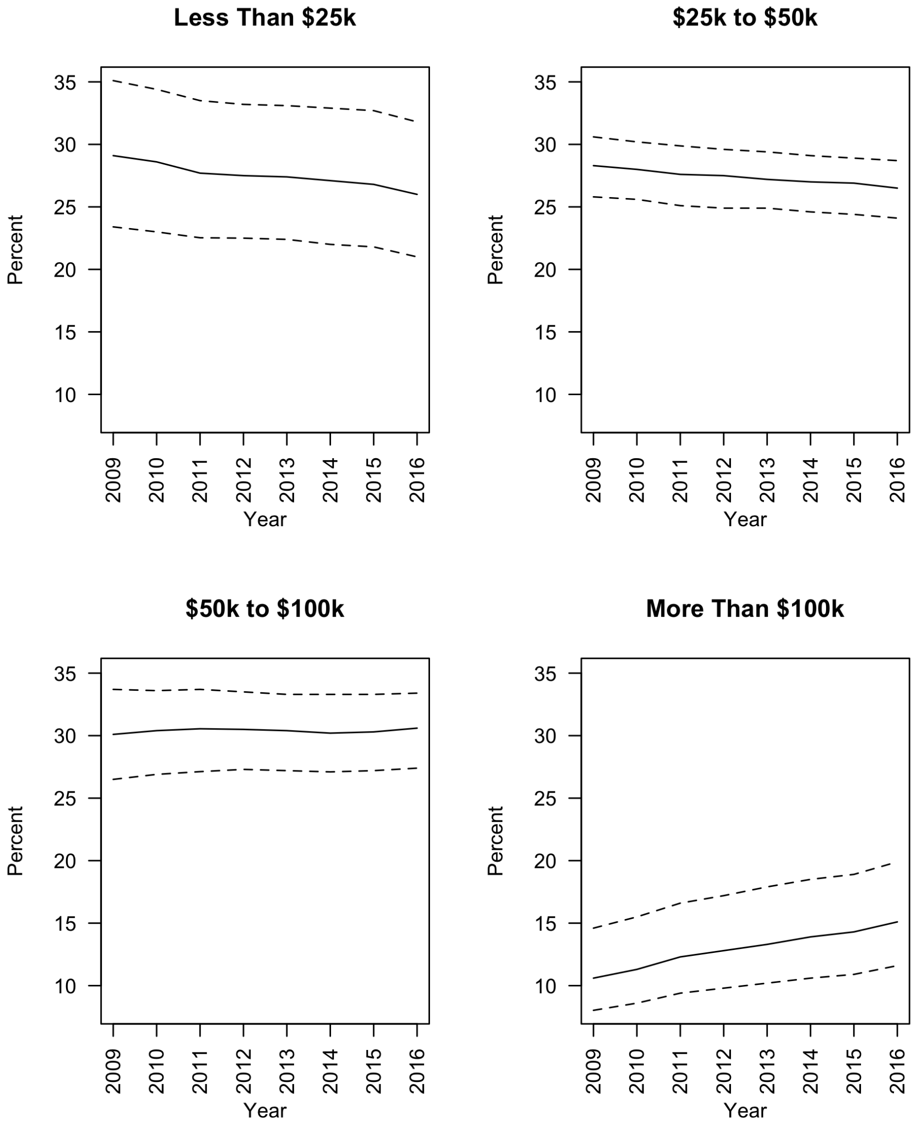
While the highest bracket is the smallest on average, it is not evenly distributed across the country. As seen below, there clear pockets of high earners, centered on major cities.

Figure 3.11: Percentage of People Earning Over $100k Annually (2016)



Though the proportion of people in the last bracket is distinctly lower than the others, it is also the only group with net growth. The average percentage of people earning $100,000 or more annually grows five percentage points from 2009 to 2016, while the average levels of people earning less than $50,000 are dropping. This does not mean however that they are moving to higher income brackets, and instead may simple be dropping out of the labor force entirely.

Figure 3.12: Inter-Quartile Ranges of Income



3.2.1 Principal Component Analysis

Across four income brackets, two clear patterns emerge. First, counties tend to have more people that are high income or that are low income. This alone explains 68% of all variation, showing it to be a major distinction between different counties. The second strongest trend found is that counties have more equitable income distributions on the one hand, versus more unequal on the other. This accounts for another fifth of variation across the four variables.

Table 3.4: Income: Principal Component Loadings (2016)

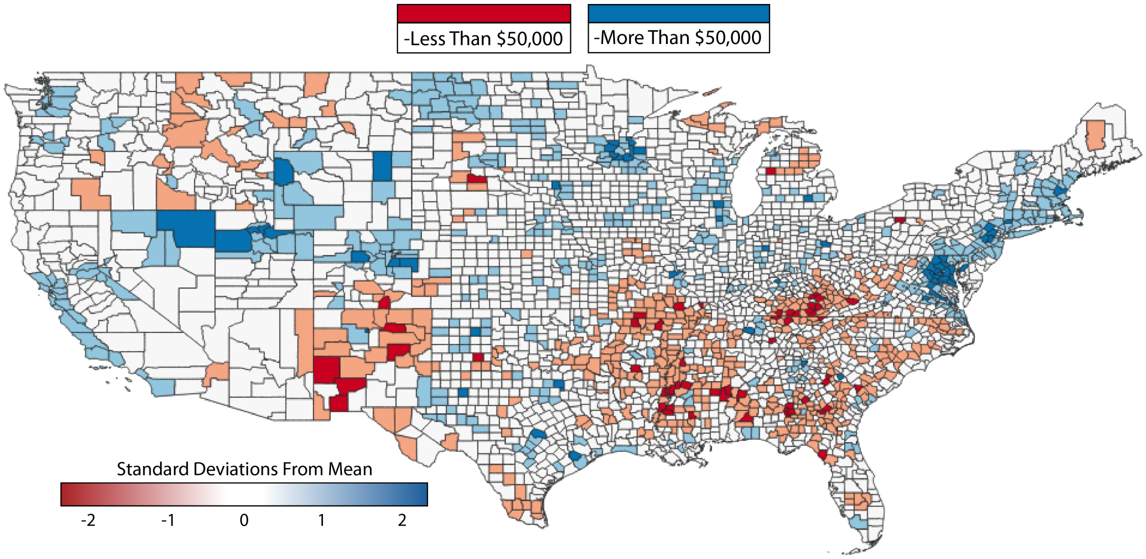
	Comp.1	**Comp.2**
Less Than $25k	-0.561	0.271
$25k to $50k	-0.417	-0.703
$50k to $100k	0.447	-0.613
More Than $100k	0.558	0.238

Table 3.5: Income: Principal Component Variances (2016)

	Comp.1	**Comp.2**
Standard deviation	1.6524	0.9225
Proportion of Variance	0.683	0.213
Cumulative Proportion	0.683	0.896

The clearest concentration of annual incomes under $50k is in the South, stretching from Arkansas down along the Gulf coast to northern Florida. Appalachia and New Mexico also have high proportions of people in these lower two income brackets. High earners are concentrated largely in major cities, though there are some interesting exceptions, particularly in north eastern Nevada, the lower central coastal counties in California, much of Wyoming, and in north western North Dakota. These do not correlate to major cities.

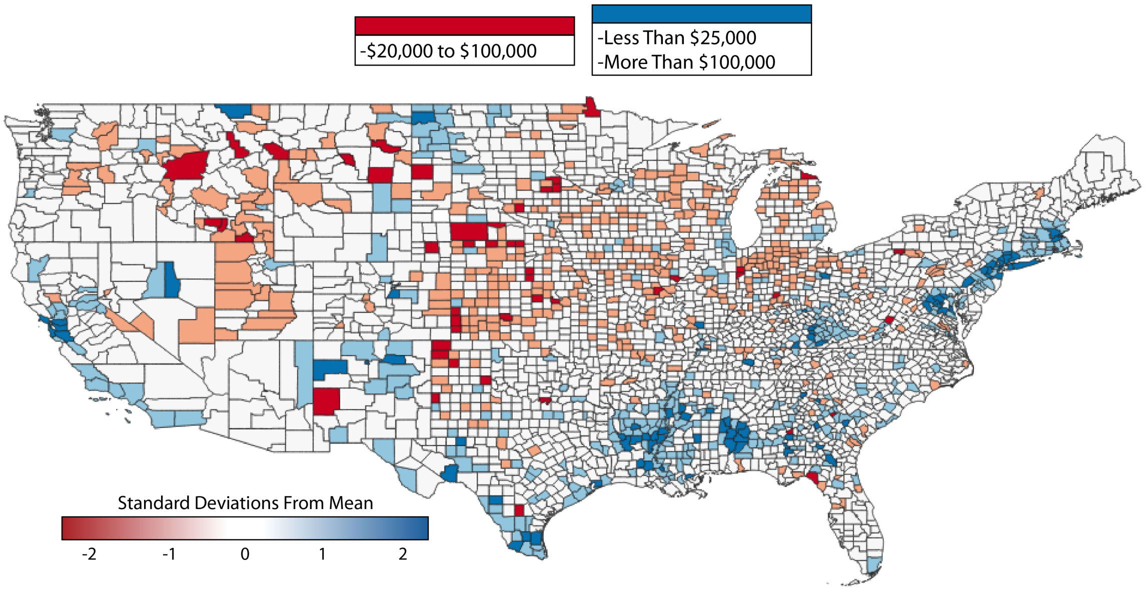
Figure 3.13: Income: Component 1 (2016)



The second component shows unequal distributions of income in blue, and more equal ones in red. The San Francisco Bay Area and the Washington D.C./Philadelphia/New York City/Boston metropolitan corridor are clear concentrations of the former phenomenon. Areas in the South, for example along the Mississippi River in Louisiana, and in southern Alabama also show a tendency towards unequal earnings as a second order effect. Finally, the counties in North Dakota that were found to have significant proportions of high earners are also found to not have equal distribution of those earnings.

The central Midwest, as well as Utah, Idaho, and parts of Montana go the other direction, with more people earning in the middle of the income spectrum.

Figure 3.14: Income: Component 2 (2016)



Loadings for Component 1 are robust with respect to time, but Component 2 is not clearly so. The trend for the middle two brackets is clear, but the opposing loadings for this component seem to be converging to zero. This means that Component 2 is predominantly picking up the presence or absence of the middle two income brackets alone, unless strong scores in the positive direction are found.

Figure 3.15: Income - Component 1: 2009 to 2016

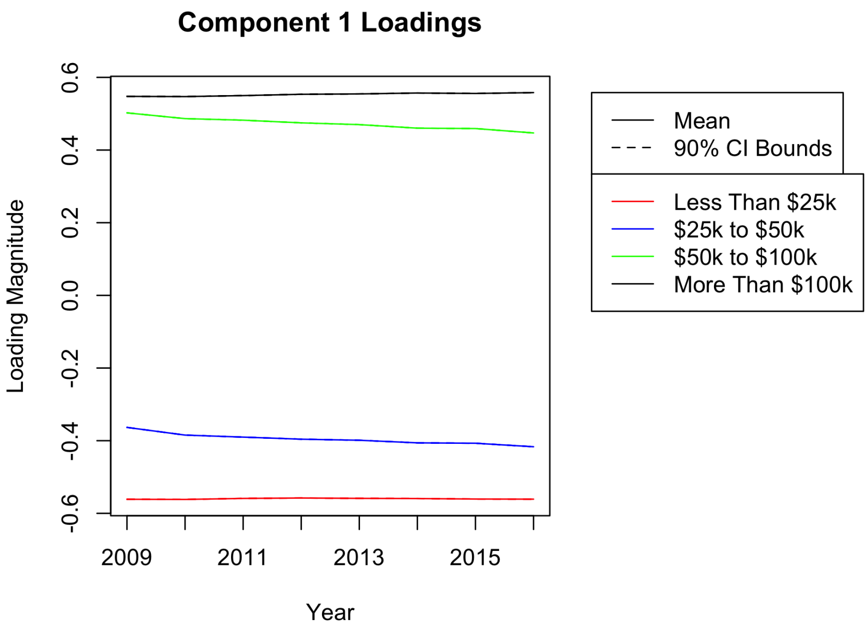


Figure 3.16: Income - Component 2: 2009 to 2016

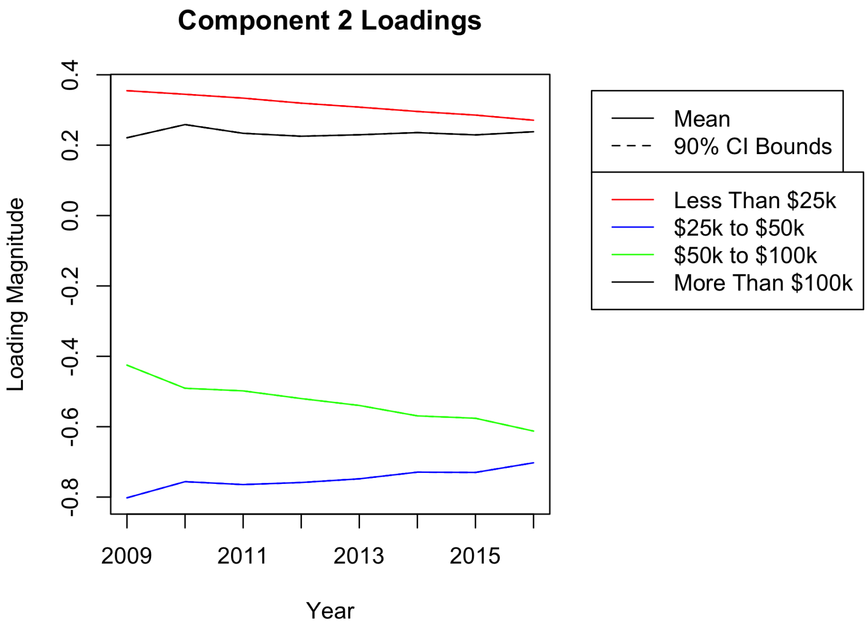
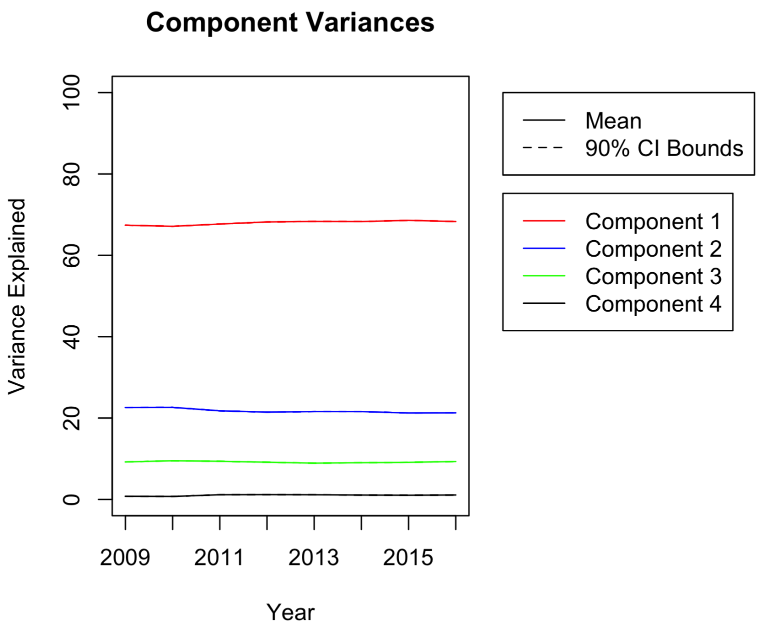


Figure 3.17: Income - Proportion of Variance Explained: 2009 to 2016



3.3 Educational Attainment

Educational attainment is the final degree a person has received at the time the survey was conducted. This data is only for people over 25, as 18-25 year olds are more likely to still be in the process of gaining their final degrees. People over 25 may still be pursing their education, but the Census uses this threshold to approximate the distribution of terminal degrees.

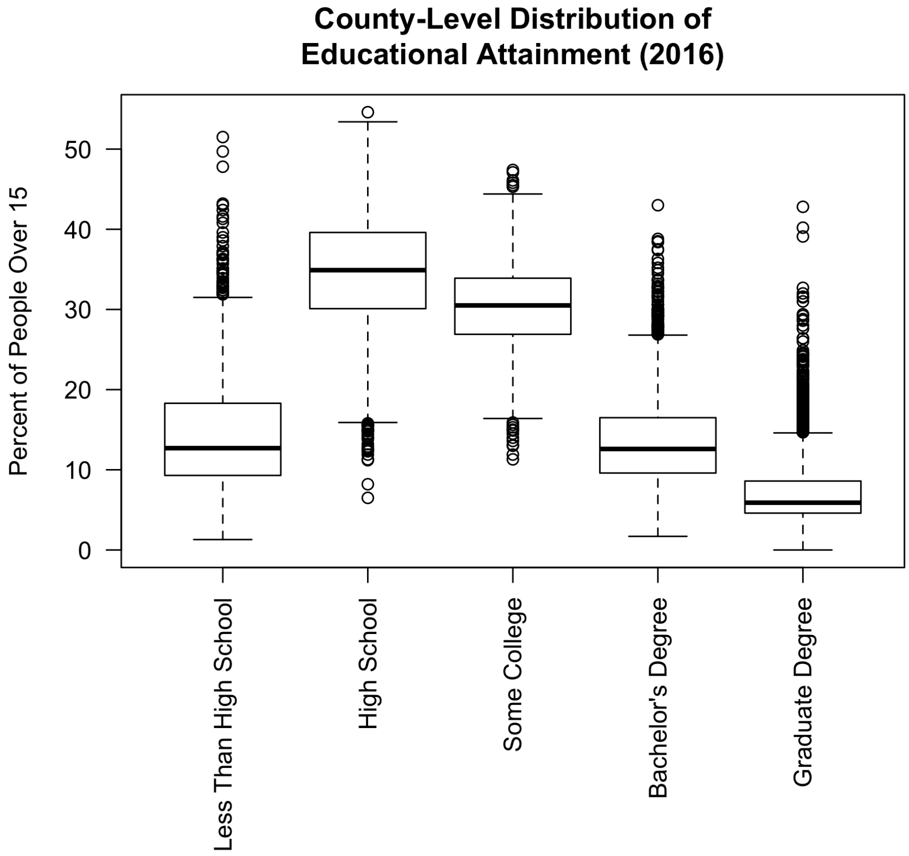
The ACS breaks educational attainment into seven categories. Preliminary principal component analysis indicates that 'Less than 9th grade' and '9th to 12th grade, no diploma' have similar behaviors. This is true for 'Some college, no degree' and 'Associate's degree' as well. Combining these into variables for 'Less than high school' and 'Some college' increases the strength of their effects. These two will be used instead of the original four for all analysis.

Table 3.6: Educational Attainment

Less than high school (Less than 9th grade, 9th to 12th grade, no diploma)
High school graduate (includes equivalency)
Some college (Some college, no degree, Associate's degree)
Bachelor's degree
Graduate or professional degree

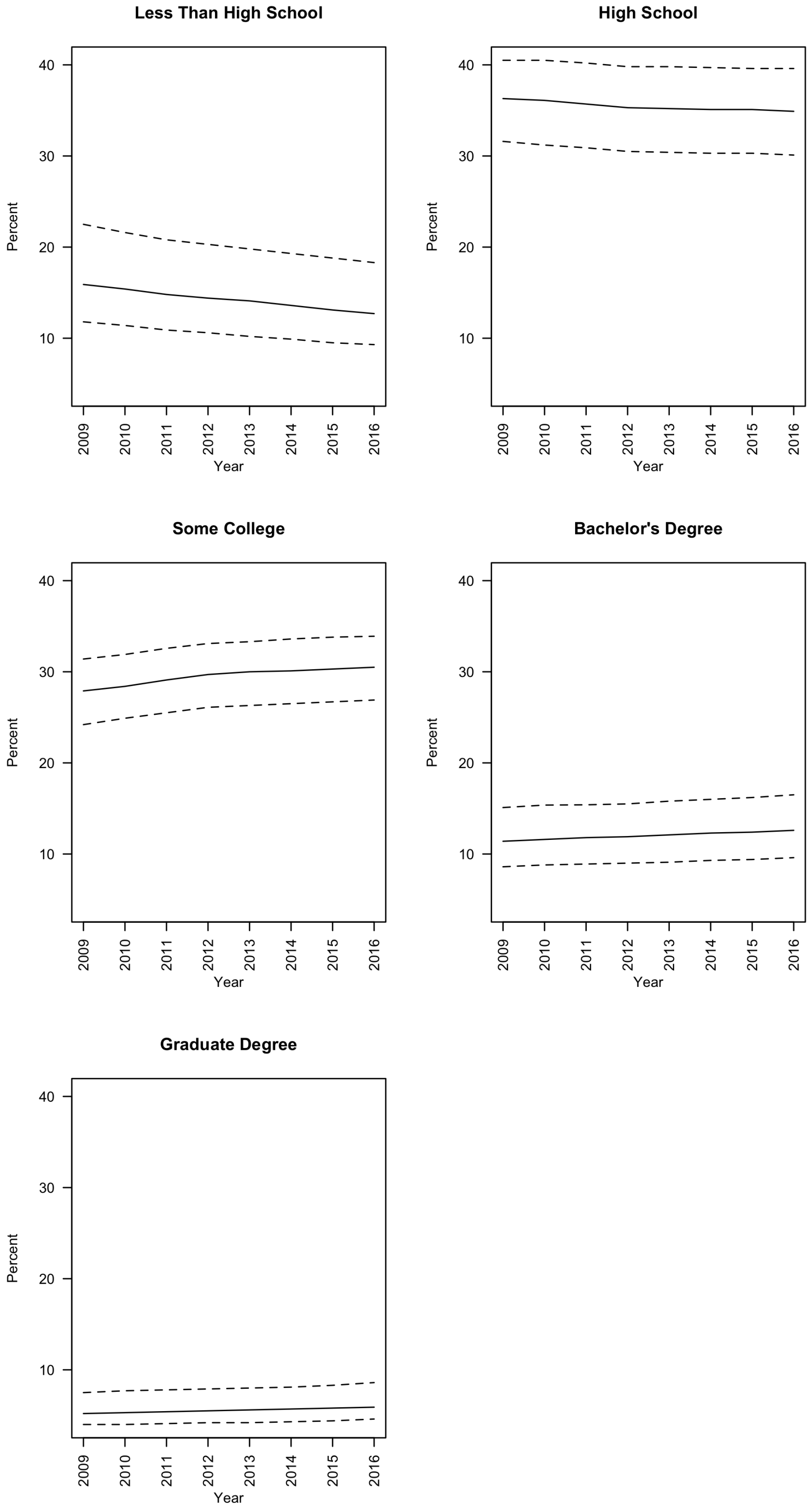
On average, a high school education is the most common terminal level for people over 25, followed closely by some college or an Associate's degree. County-level averages for people with less than a high school education, and with a Bachelor's degree are similar, with higher degrees making up consistently less than 10% of the total population. In spite of this low average, there are some outliers with seemingly very high percentages of people having graduate or professional degrees. The highest is Falls Church City, VA. Falls Church is an independent city with a population of only around 14,500 people, and is almost directly across the Potomac River from Washington D.C. The county with the second highest proportion of graduate degrees is Los Alamos County, home to the Los Alamos National Labratory, and with a total population of just over 18,000. The third outlier is Arlington County, another suburb of Washington, D.C. This county has a population of almost 235,000 people, making the concentration of people with graduate educations truly striking.

Figure 3.18: Distribution of Educational Attainment (2016)



Looking at these levels over time, we can see that the percent of the population with less than a high school education is in clear decline. This is true for the percent of people with only a high school education as well, though to a lesser extent. Correspondingly, the various levels of college education are increasing, though each higher level is increasing by a slower rate than the previous over the study period.

Figure 3.19: Inter-Quartile Ranges of Educational Attainment



3.3.1 Principal Component Analysis

The first trend found in this data is that counties tend to group by more education or less education. This distinction alone accounts for almost 60% of variation in the data. The second strongest pattern found is between counties with high proportions of people with some college education, versus a bimodal distribution where people tend to have either graduate or professional degrees, or less than a high school education. This component contains 22% of total variation.

Table 3.7: Educational Attainment: Principal Component Loadings (2016)

	Comp.1	**Comp.2**
Less Than High School	-0.456	0.310
High School	-0.458	
Some College	0.252	-0.815
Bachelor's Degree	0.544	
Graduate Degree	0.472	0.449

Table 3.8: Educational Attainment: Principal Component Variances (2016)

	Comp.1	**Comp.2**
Standard deviation	1.7165	1.0506
Proportion of Variance	0.589	0.221
Cumulative Proportion	0.589	0.810

Educational attainment is clearly not evenly distributed across the country. Counties with more of a trend towards a high school education or less are concentrated in the South and East, whereas at least some college is more strongly associated with big cities across the country. For the second component, a striking concentration of counties in the upper and central Midwest have higher proportions of people with some college experience, as do many counties in the West. Counties with an unequal distribution of education are concentrated in the big cities on the coasts, notably the

San Francisco Bay Area, Washington, D.C., and the New York City/Philadelphia area.

Figure 3.20: Educational Attainment: Component 1 (2016)

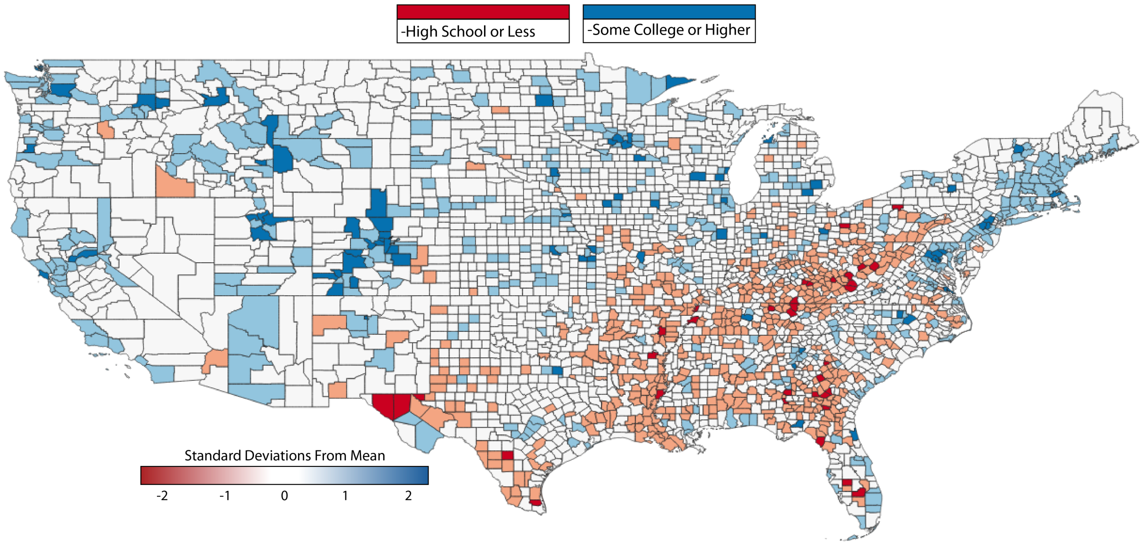
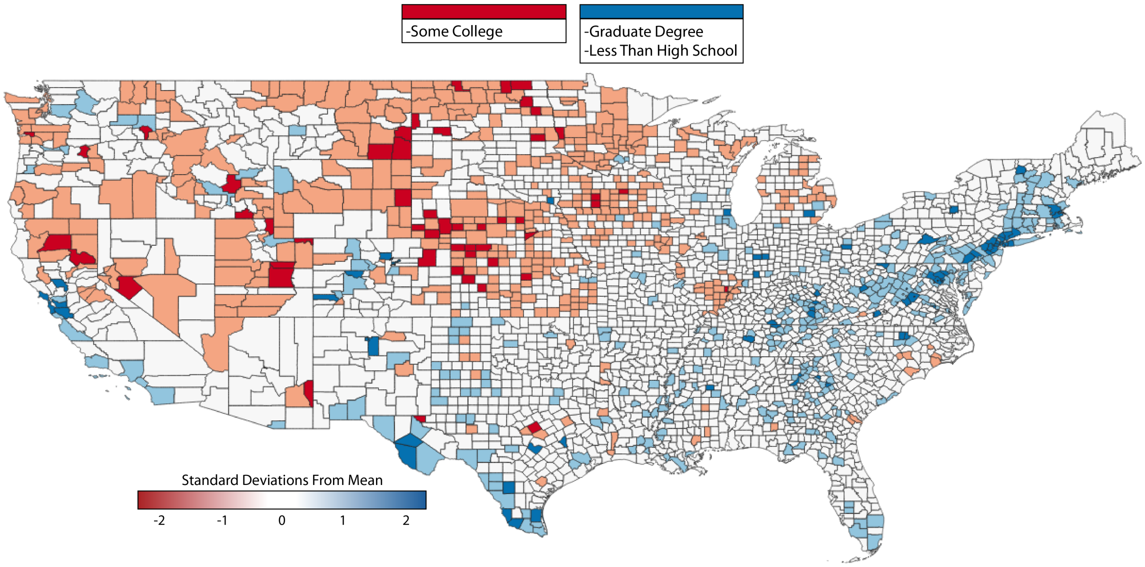


Figure 3.21: Educational Attainment: Component 2 (2016)



These components are mostly stable over time. People with some college education is becoming a less useful indicator for if a county falls into the 'high education' category in Component 1, even as Bachelor's and graduate degrees become more important. For the second component, more people with a high school education was originally more closely associated with some college education, but over the study period these two indicators have diverged in their ability to describe the educational distribution of a county.

Figure 3.22: Educational Attainment - Component 1: 2009 to 2016

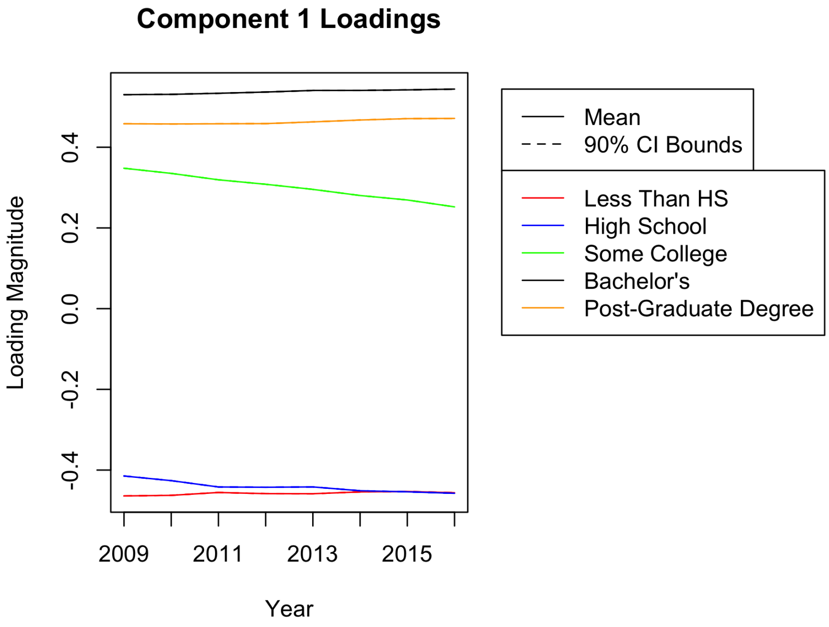


Figure 3.23: Educational Attainment - Component 2: 2009 to 2016

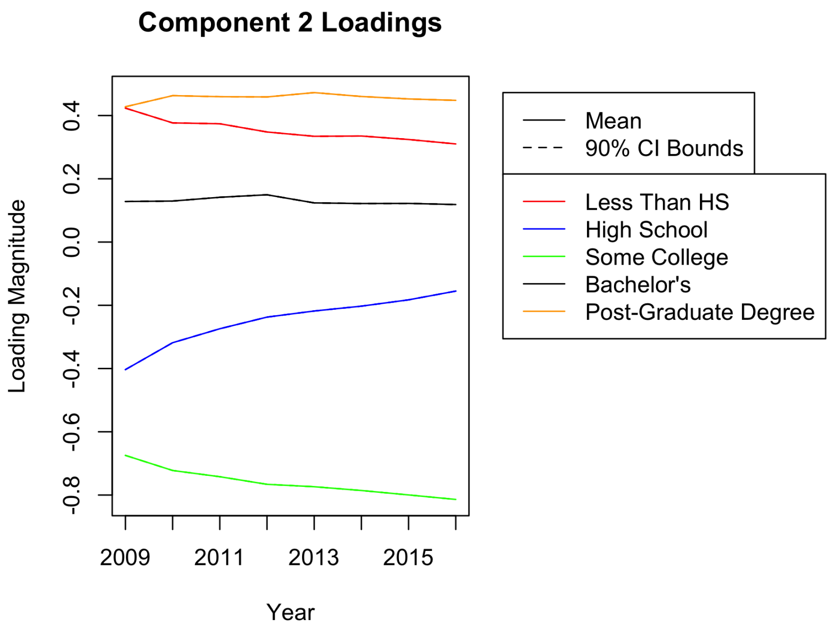
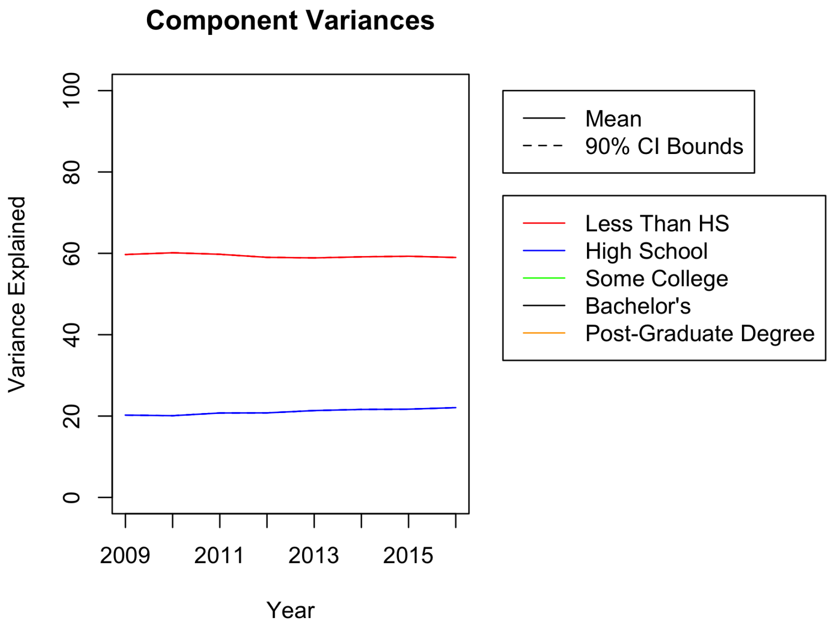


Figure 3.24: Educational Attainment - Proportion of Variance Explained: 2009 to 2016



3.4 Field of First Bachelor's Degree

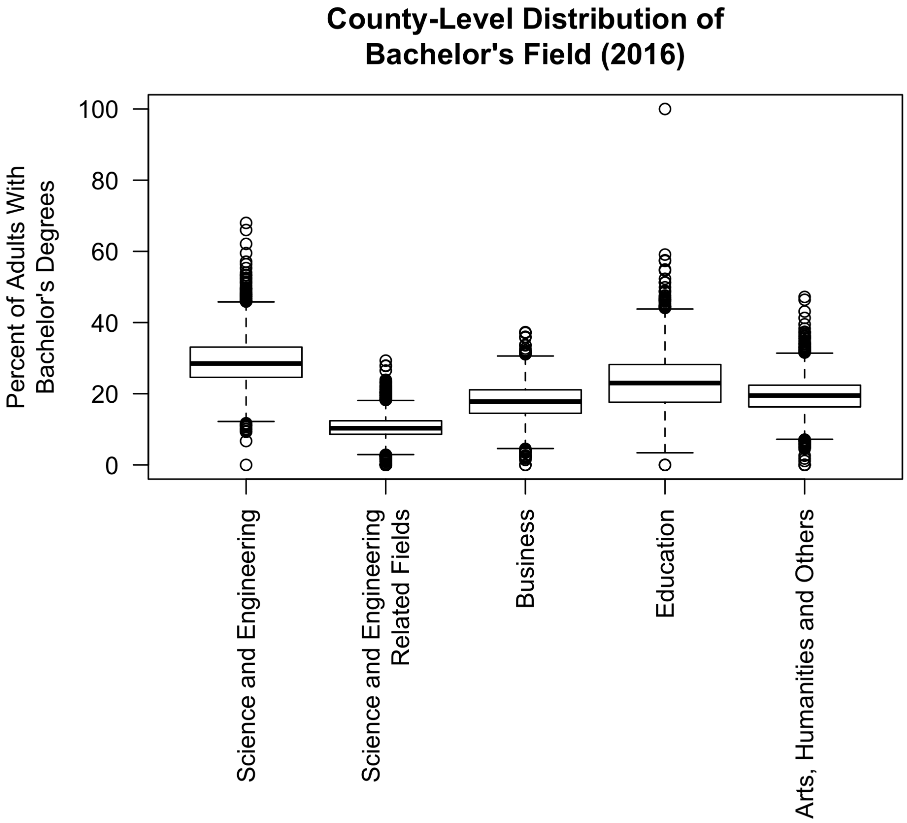
Bachelor's degrees are grouped by the Census into five categories. These are all widely recognized areas except possibly for those categorized as 'Science and Engineering Related'. Degrees in this category are ones that are technical, but not explicitly in scientific or engineering fields. Examples are degrees in computer science, nursing, or science teacher education [2]. In the fifth category of degree fields, 'Other' includes fields like public administration, criminal justice, and social work. Unlike most other analyses herein, this data is only available for 2015 to 2016.

Table 3.9: Bachelor's Degree Fields

Science and Engineering	Science and Engineering Related Fields
Business	Education
Arts, Humanities and Others	

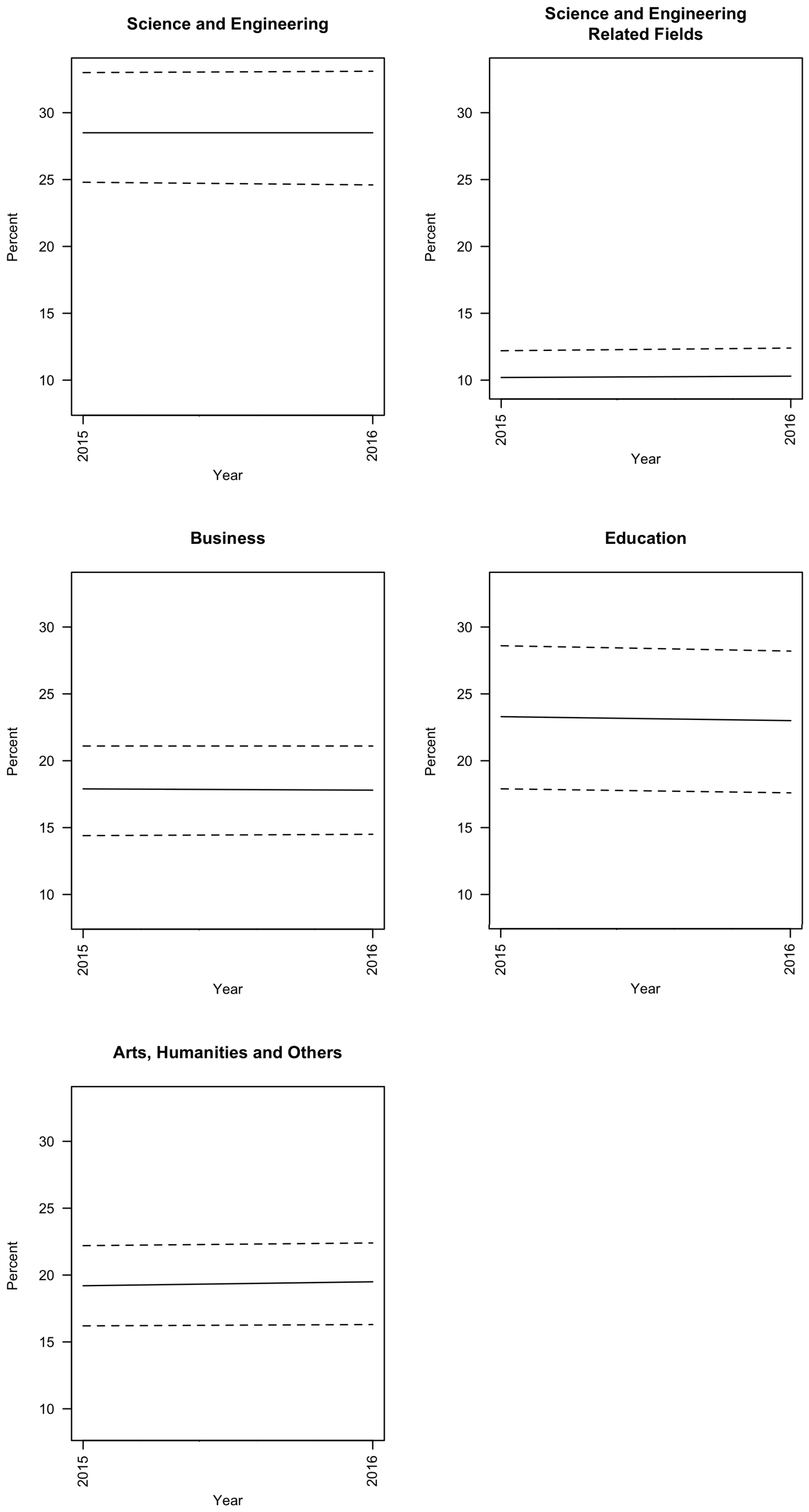
Science and engineering degrees are on average the most common, followed by education. Business degrees and degrees in the arts, humanities, and other fields are on average not much below these first two, but science and engineering related fields are distinctly less common than the other four degree categories. Loving County, Texas is the outlier for education degrees, where all adults with Bachelor's degrees have their first degree in education. Loving County is the least populous county in the country, with just 134 people, so it is possible that the only people with Bachelor's degrees are the local teachers.

Figure 3.25: Distribution of First Bachelor's Degree Field (2016)



Degrees by field do not show any major trends over the study period, though education degrees as a proportion of all degrees seem to be slowly declining.

Figure 3.26: Inter-Quartile Ranges of Bachelor's Field



3.4.1 Principal Component Analysis

The most important distinction between counties found is between science, engineering, arts, humanities, and other degrees on the one hand, versus education degrees plus science and engineering related fields on the other. This second group can be though of as technical non-scientific degrees, and their association here is plausible in this respect. This component accounts for just over 40% of total variation. The second component finds that counties differentiate by having more business degrees, versus more in science, engineering, and education degrees. The third trend find a distinction between more education degrees in a county, versus more science and engineering related fields, as well as more arts and humanities degrees. Finally, the fourth component shows that counties can tend to have more science and engineering degrees, versus more arts and humanities degrees. These last three components account for between 16% and 23% of all variation, which is relatively strong for lower comoponents. This signifies that each of these trends is clearly detectable to the principal component analysis technique.

Table 3.10: Bachelor's Field: Principal Component Loadings (2016)

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**
Science and Engineering	0.539	-0.284		0.597
Science and Engineering Related Fields	-0.345		0.884	
Business		0.906		
Education	-0.600	-0.309	-0.346	
Arts, Humanities and Others	0.480		0.249	-0.744

Table 3.11: Bachelor's Field: Principal Component Variances (2016)

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**
Standard deviation	1.4416	1.0765	0.9574	0.9137
Proportion of Variance	0.416	0.232	0.183	0.167
Cumulative Proportion	0.416	0.648	0.831	0.998

Interestingly the first component shows a clear concentration of science, engineering, arts, and humanities degrees on the East and West coasts. Much of Colorado shares this pattern as well. Component 2 finds high proportions of business degrees across the South, with education and more technical degrees in scattered pockets across the West.

The only major geographic concentration of scores for the third component finds high proportions of education degrees in western Texas and New Mexico. With the relatively rural nature of these counties, this likely reflects a dearth of Bachelor's degrees in general. This would mean that it is mostly grade school teachers that possess this level of education or higher. Lastly, the fourth component is significant statistically, but is not geographically concentrated in any meaningful way, as shown in the final map.

Figure 3.27: Bachelor's Field: Component 1 (2016)

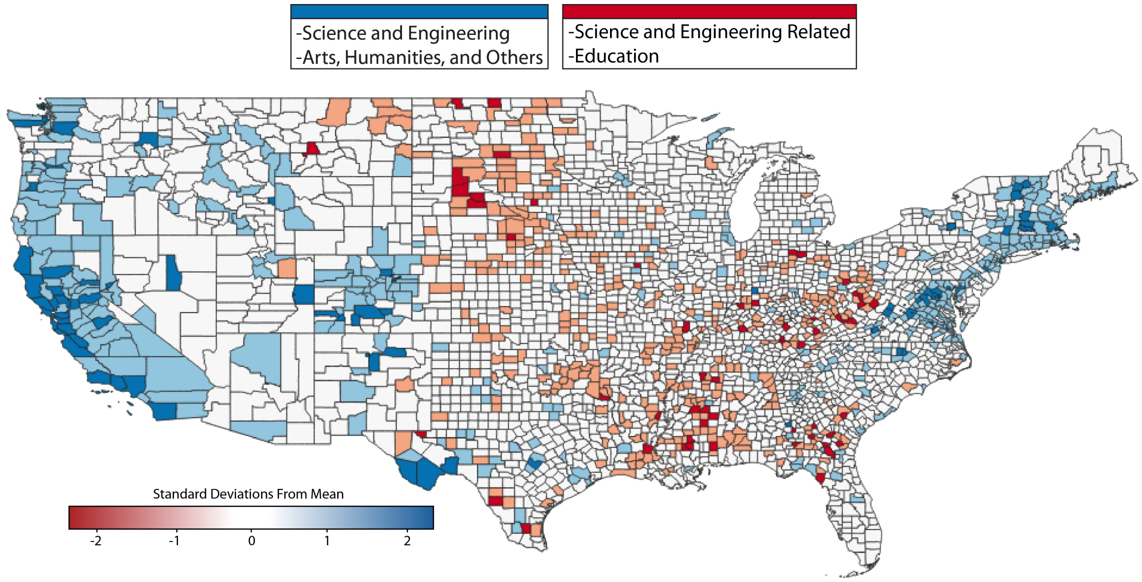

-Science and Engineering
-Arts, Humanities, and Others
-Science and Engineering Related
-Education
Standard Deviations From Mean
-2
-1
0
1
2

Figure 3.28: Bachelor's Field: Component 2 (2016)

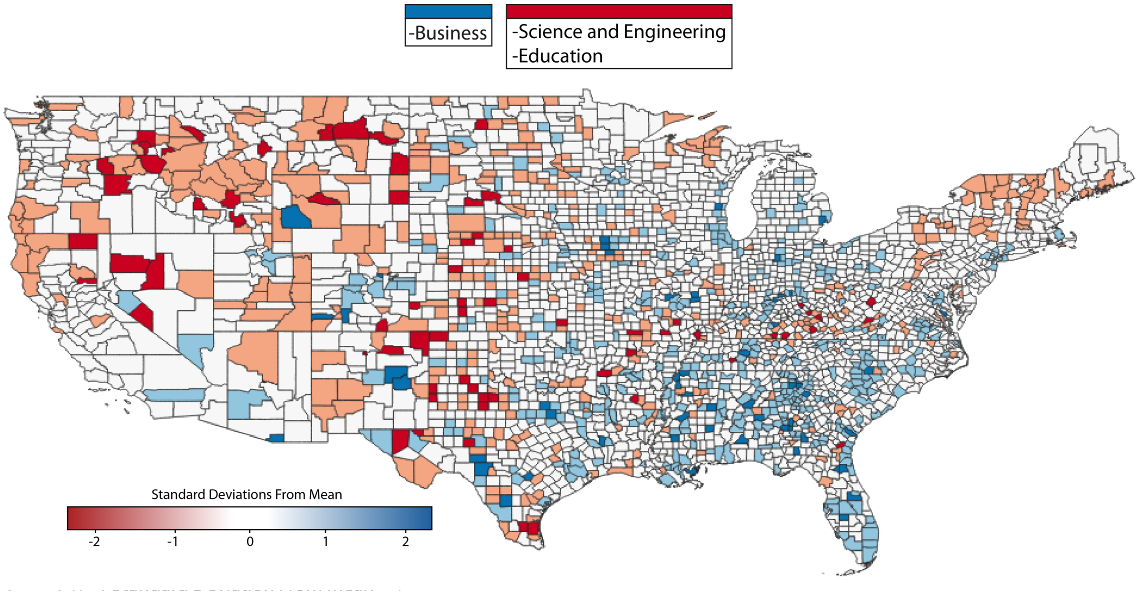


Figure 3.29: Bachelor's Field: Component 3 (2016)

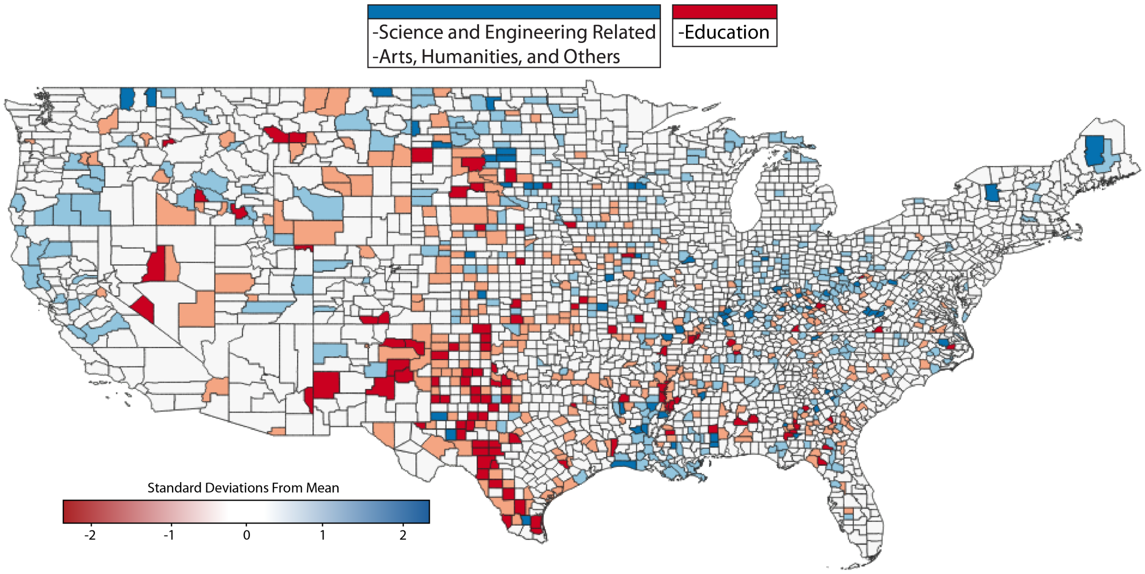
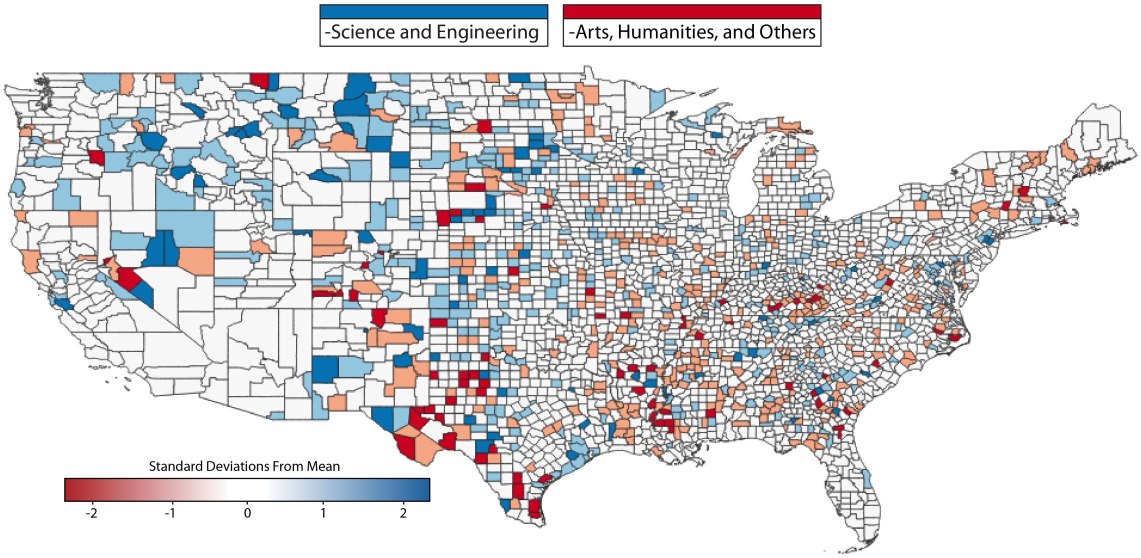

-Science and Engineering Related
-Arts, Humanities, and Others
-Education
Standard Deviations From Mean
-2
-1
0
1
2

Figure 3.30: Bachelor's Field: Component 4 (2016)



There are some fluctuations in the relative loadings describing the latent structure of these variables, but with only two years of data, it is less clear that these are important trends. Even in the later components, there is very little re-ordering of variables, and the variances are almost completely unchanged.

Figure 3.31: Bachelor's Field - Component 1: 2015 to 2016

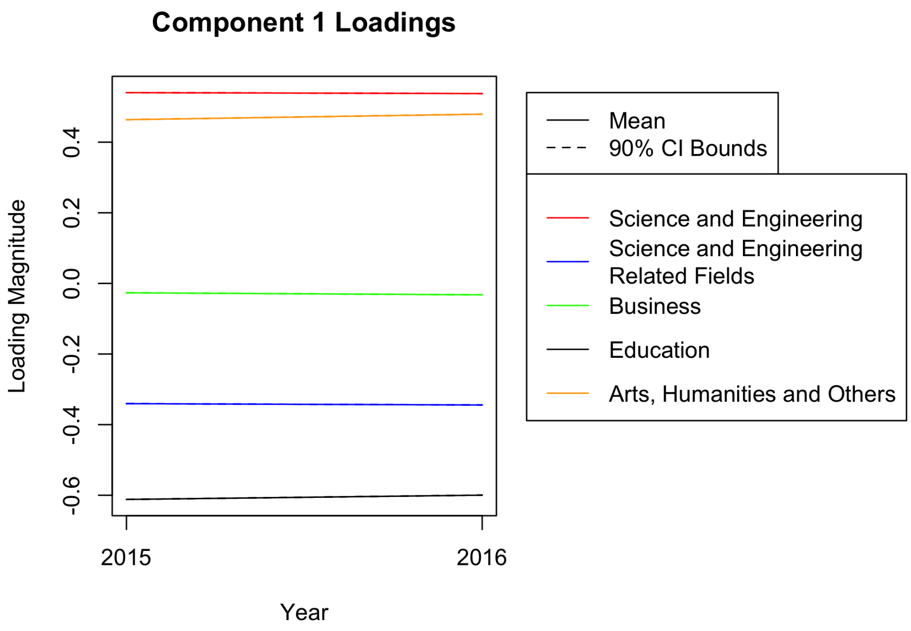


Figure 3.32: Bachelor's Field - Component 2: 2015 to 2016

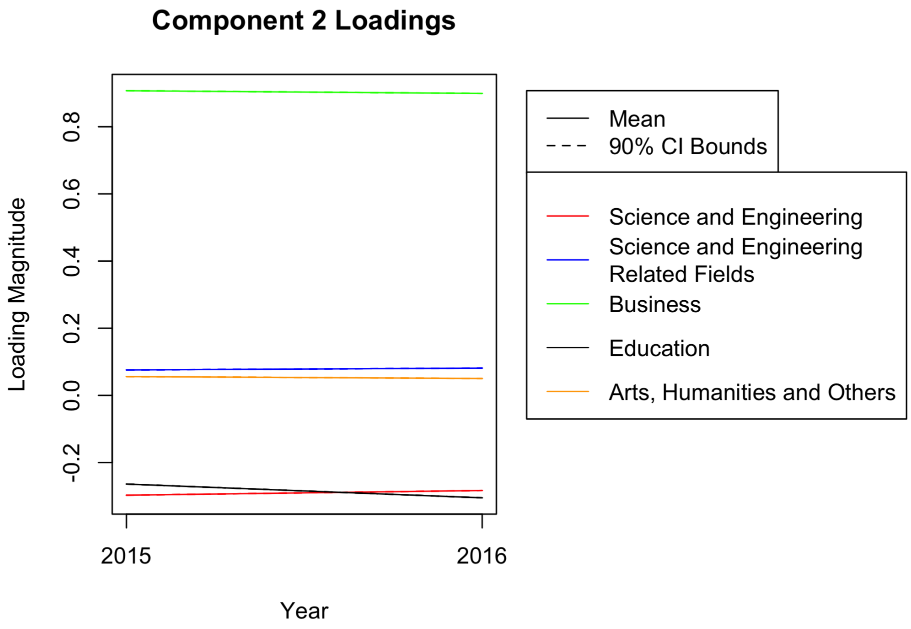


Figure 3.33: Bachelor's Field - Component 3: 2015 to 2016

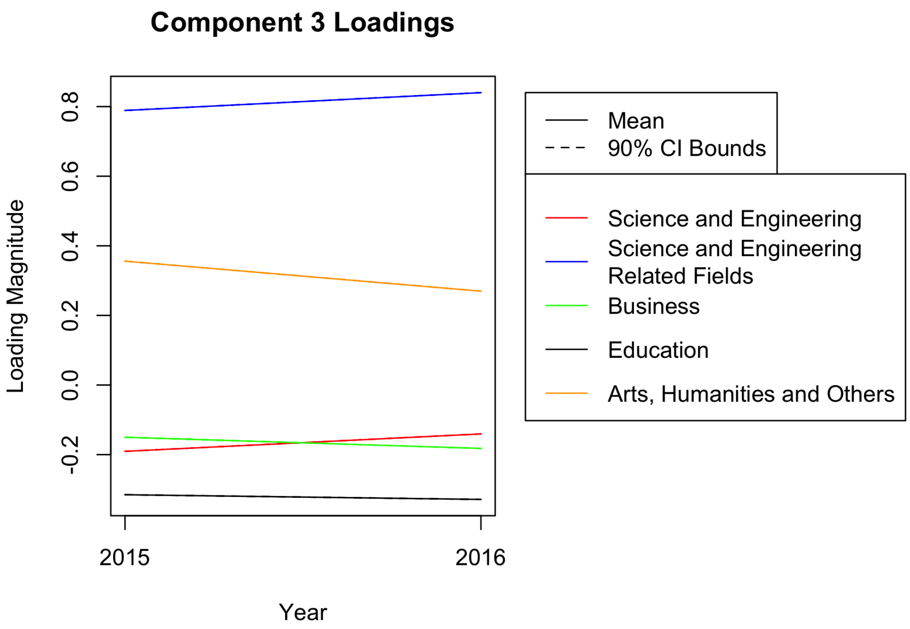


Figure 3.34: Component 4: 2015 to 2016

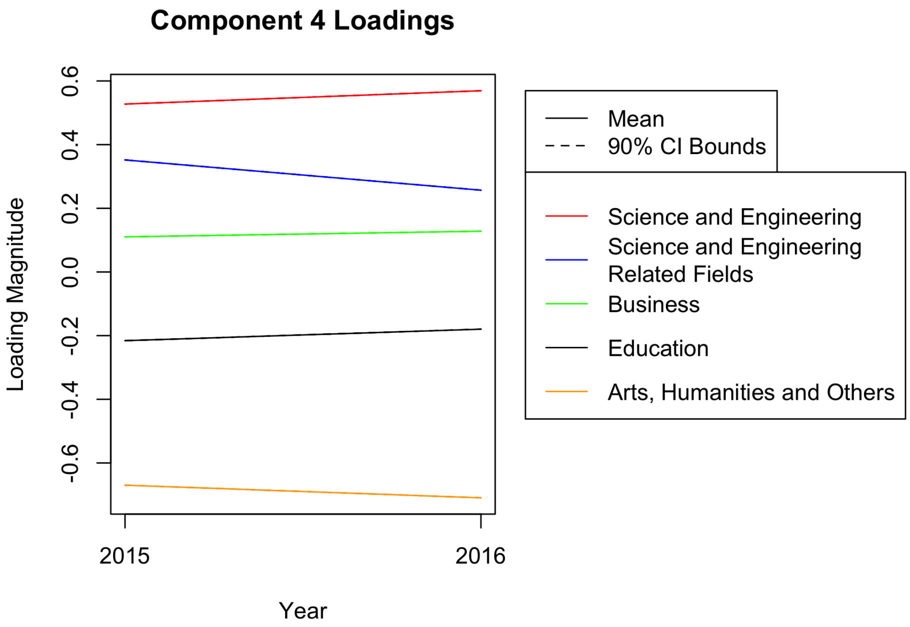
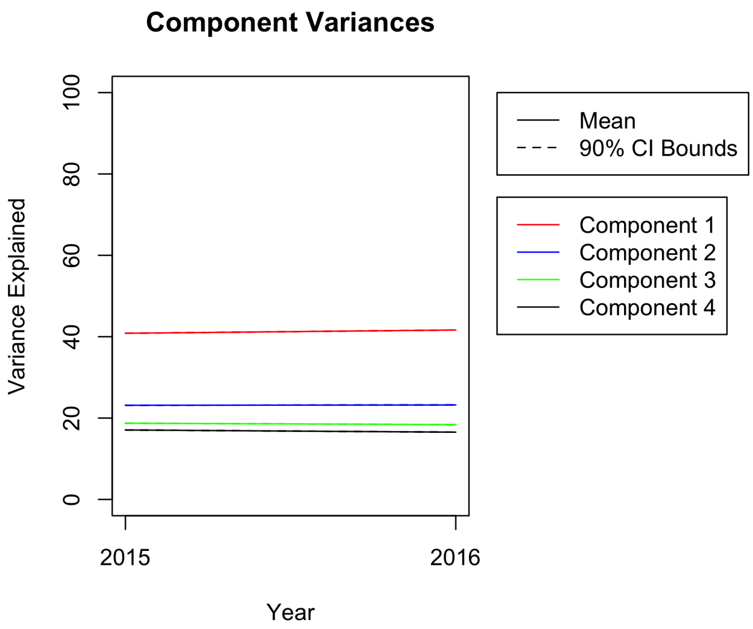


Figure 3.35: Bachelor's Field - Proportion of Variance Explained: 2015 to 2016



3.5 Commuting Characteristics: Places of Residence and Employment

It is not uncommon for people to live away from their places of employment. What types of political and social borders are crossed in daily commutes is examined here at the city and county levels. For cities we use Census designated Places. These aim to be ”statistical counterparts of incorporated places” [3], and include un-incorporated areas if they contain notable concentrations of people. All data in this section are for workers over the age of 16.

In this data, there are two groups with three categories each. Each group covers 100% of the population in each county.

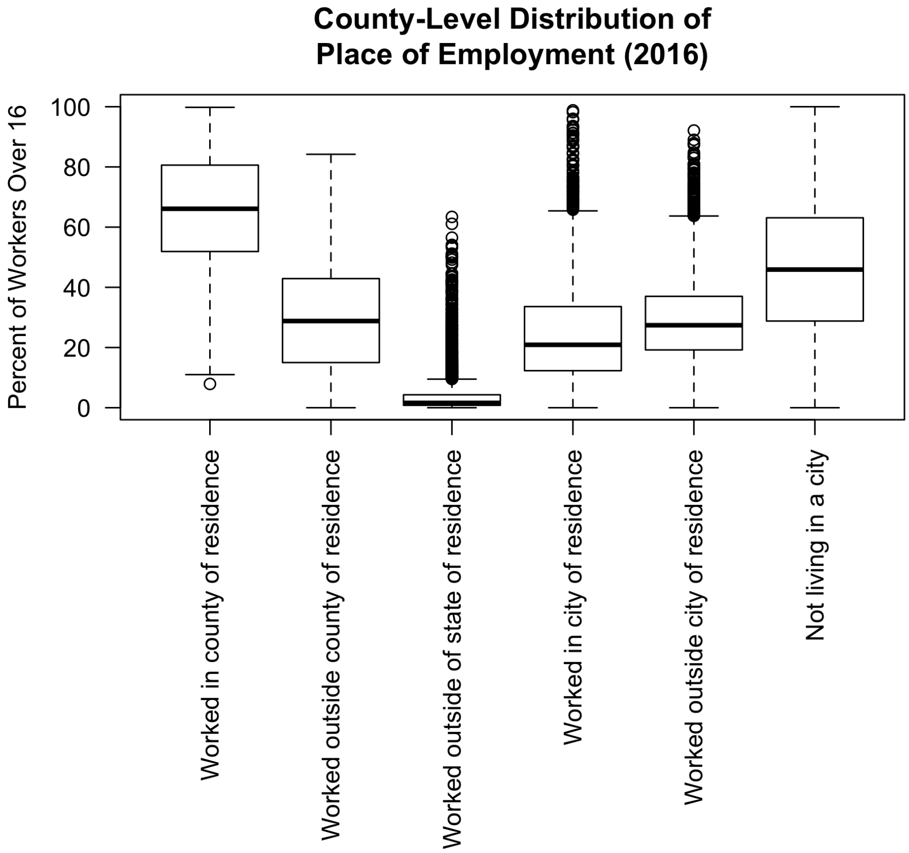
Table 3.12: Places of Residence and Employment

County-Level	**City-Level**
Worked in county of residence	Worked in city of residence
Worked outside county of residence	Worked outside city of residence
Worked outside of state of residence	Not living in a city

On average, at the county level, most people live in the same county as they work, shown in Table 3.36. Approximately 25%-45% of people live in the same state, but not the same county as their job, and on average, less than 10% of people cross state lines in their commutes.

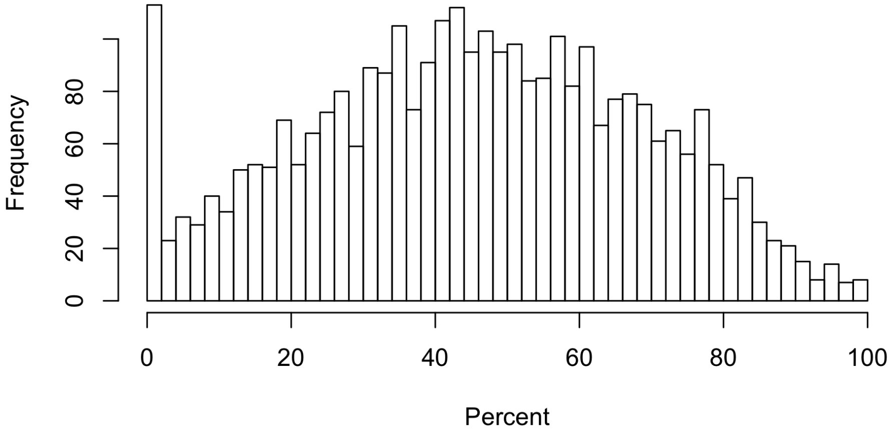
Looking at the city-level (Census Place), it is much more common for Americans to cross city borders during their commute than county borders. However, there is not much difference between average percentages of people who work within their home cities, versus those who commute across city borders. This is likely because a Census Place is often part of a larger urban area. For example, there are 134 designated cities in the San Francisco-Oakland urban area, and moving between them might not be meaningful to people living and working within this wider urban agglomeration.

Figure 3.36: Place of Employment and Residence Distribution across U.S. Counties



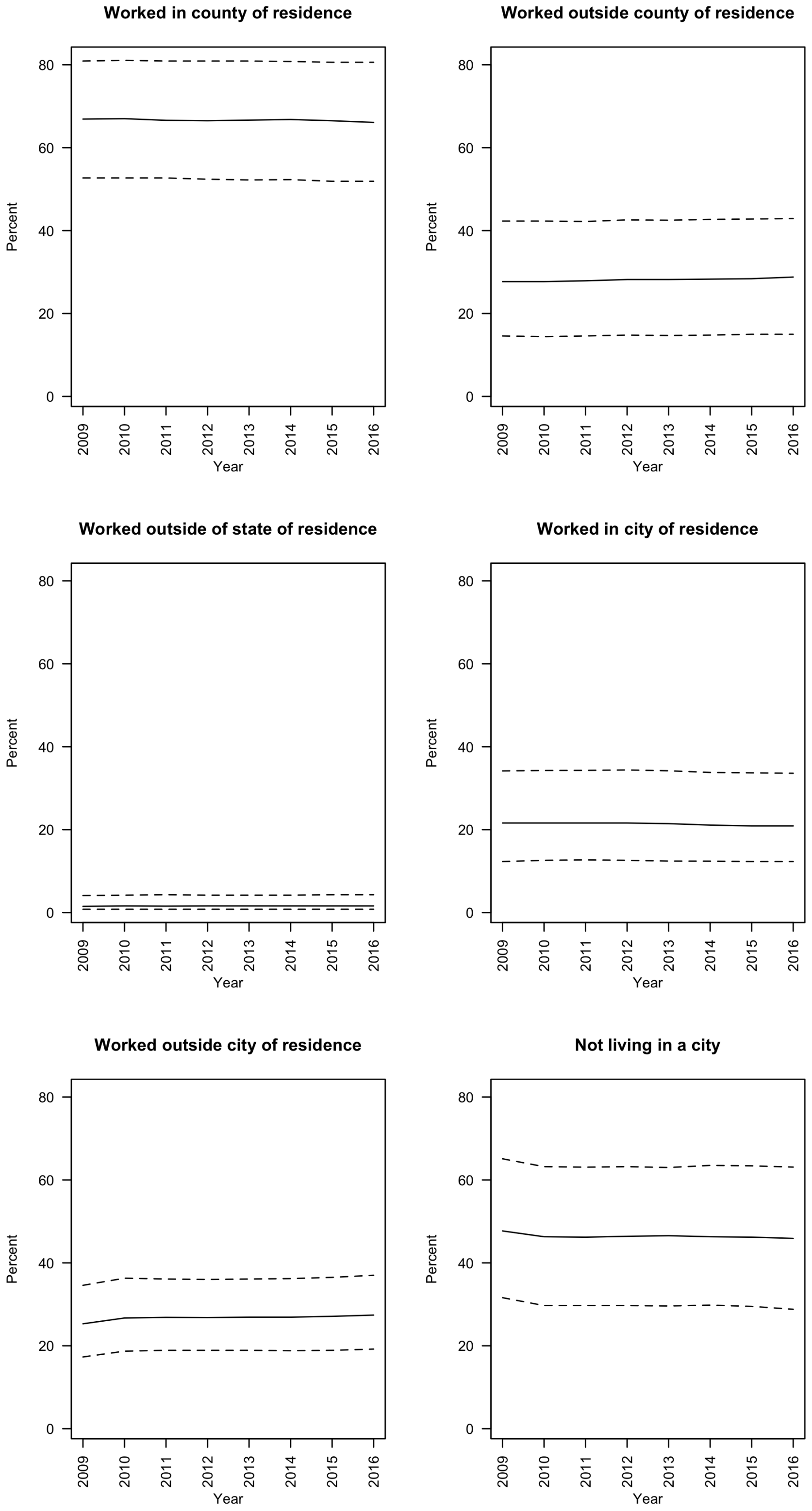
Interestingly, people who don't live in a designated city is the category with the highest average, though the distribution is clearly not concentrated at any level. There is also a clear spike at zero percent (Table 3.37), indicating counties that are entirely made up of Census designated cities.

Figure 3.37: Distribution of Workers Living Outside of City across U.S. Counties (2016)



These patterns do not change almost at all across the study period of 2009 to 2016. There is a slight uptick in the first year for the average percent of people who worked outside of their residence, and a corresponding slight decrease of people not living in a city.

Figure 3.38: Inter-Quartile Ranges of Residential and Employment Locations



3.5.1 Principal Component Analysis

The strongest trend found in this data is between people who work in the cities and counties of their residence, versus people who live outside of a city and cross county lines to get to work. This trend accounts for around half of all variation across the six variables.

The second strongest trend is between people who live outside of a city, but work in their local county in the positive direction, against people who work outside of their city and county of residence in the negative direction. Urban areas are often divided along county lines, so in this case, working outside of ones home county is likely synonymous with working outside of ones home city. Just under 25% of all variation is explained by this trend.

The third component of interest indicates the presence of high proportions of people who work outside of their state of residence. This trend alone accounts for 17% of variation across all the variables.

Table 3.13: Places of Residence and Employment: Principal Component Loadings

	Comp.1	**Comp.2**	**Comp.3**
Worked in county of residence	0.487	0.385	
Worked outside county of residence	-0.459	-0.404	
Worked outside of state of residence			0.933
Worked in city of residence	0.526		
Worked outside city of residence		-0.701	
Not living in a city	-0.467	0.435	

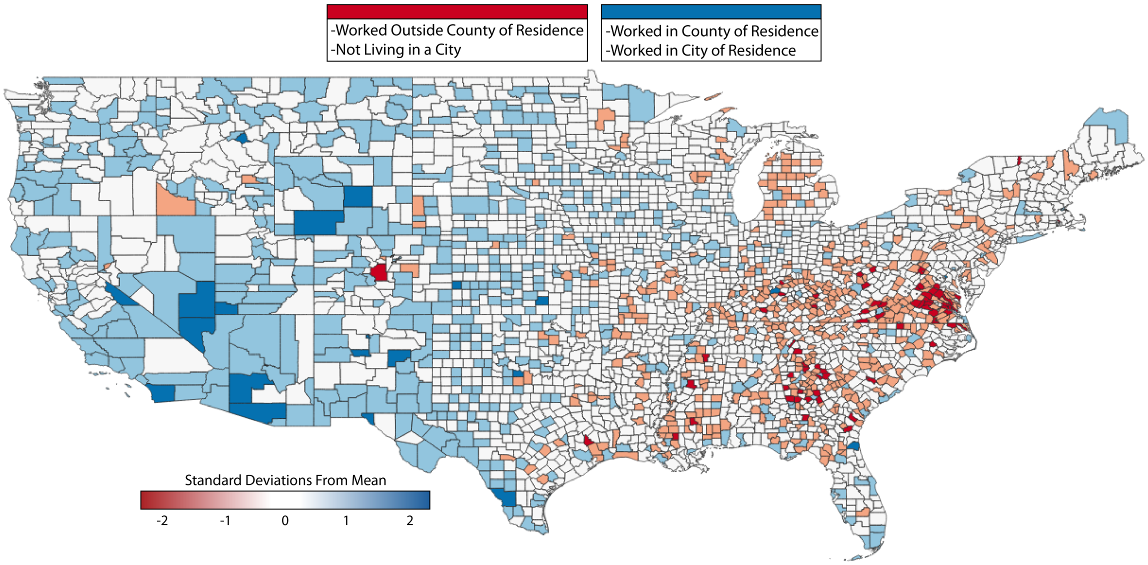
Table 3.14: Places of Residence and Employment: Principal Component Variances

	Comp.1	**Comp.2**	**Comp.3**
Standard deviation	1.7387	1.2019	1.0292
Proportion of Variance	0.504	0.241	0.177
Cumulative Proportion	0.504	0.745	0.921

When viewed geographically, the first principal component is clearly a regionally segregated phenomenon. The western half of the country has predominantly positive scores for counties, meaning higher proportions of people work in both the same city and the same county as their home. This may reflect the relatively low-density nature of the area, where commuting between small cities is less likely than in the more densely populated east.

Conversely, counties in the eastern portion of the country, and especially the South, have more people who don't live in a city and who cross county lines in their commutes. The two strongest concentrations of this effect are in the Richmond, VA/Washington, D.C. area, and the Atlanta, GA area. This may be a sign that rural counties in these areas are predominantly populated by people who commute to the nearby major urban hub. This is the most plausible explanation for the county in Colorado with a strong negative score. This is Park County, which is directly to the southwest of Denver, CO, and is therefore plausibly a county with many commuters.

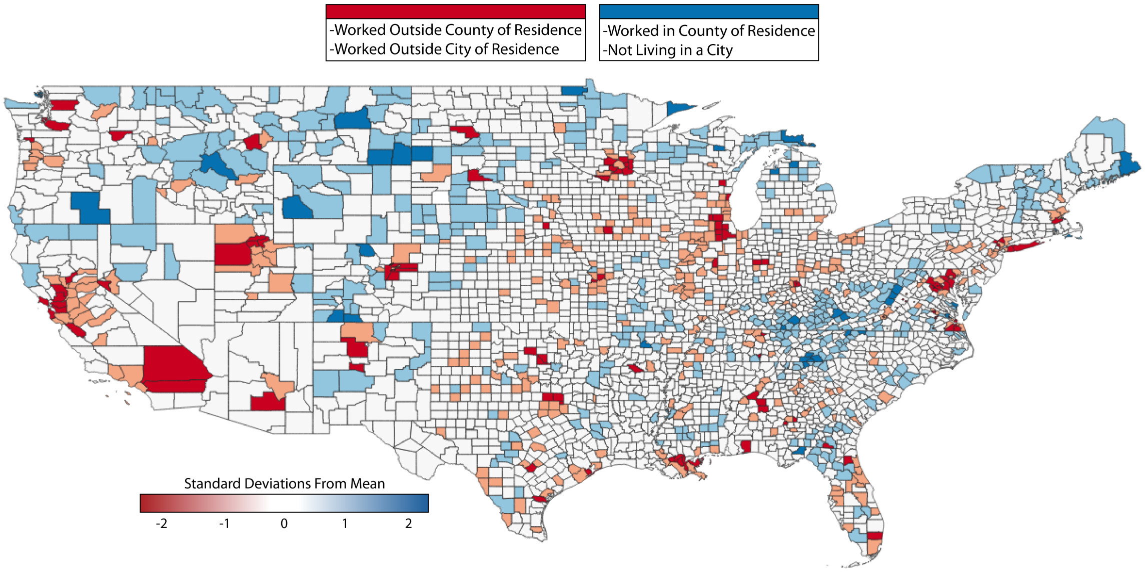
Figure 3.39: Place of Employment: Component 1 (2016)


-Worked Outside County of Residence
-Not Living in a City
-Worked in County of Residence
-Worked in City of Residence
Standard Deviations From Mean
-2
-1
0
1
2

The second component shows concentrations of people who work in their home county, but do not live in a designated city. Counties emphasizing this trend can be found in almost every state, but there are particular concentrations in the Appalachian region and in the states of Idaho, Montana, Wyoming, Colorado, Nebraska, and South Dakota.

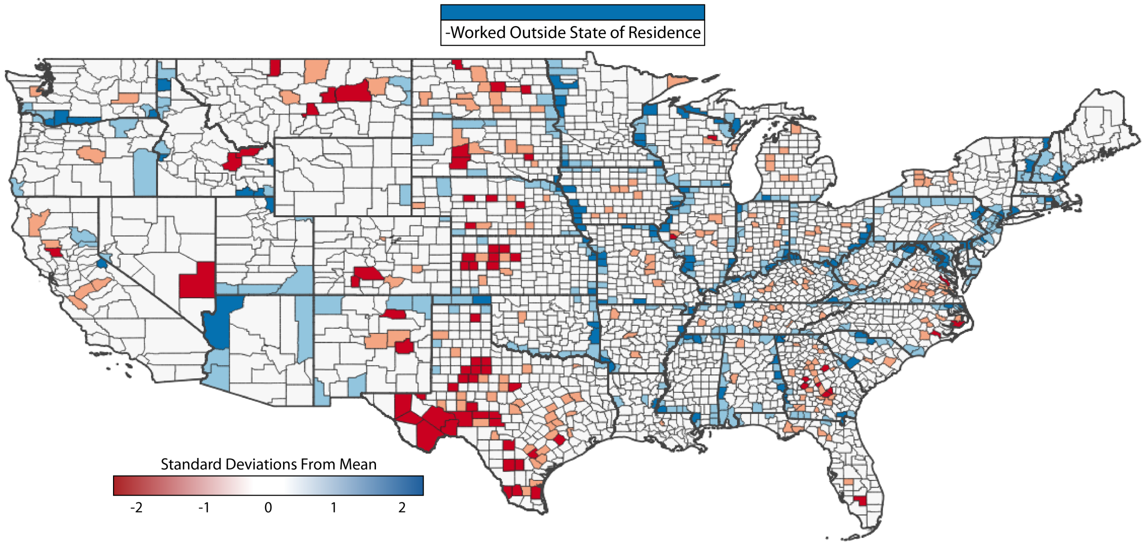
The trend in the negative direction for Component 2 is that of commuters. This becomes clear when looking at the map below, with concentrations of red surrounding every major city and state capital.

Figure 3.40: Place of Employment: Component 2 (2016)


-Worked Outside County of Residence
-Worked Outside City of Residence
-Worked in County of Residence
-Not Living in a City
Standard Deviations From Mean
-2
-1
0
1
2

The third component finds a trend of counties with high percentages of people crossing state lines to get to work. This tendency is relatively weak in the West, but the borders of many states can be clearly seen traced in blue in the eastern half of the country. Here negative scores mean a strong trend away from commutes across state lines. With few exceptions, these are understandably all away from state borders.

Figure 3.41: Place of Employment: Component 3 (2016)


-Worked Outside State of Residence
Standard Deviations From Mean
-2
-1
0
1
2

Examining the bootstrapped results across years shows that these trends and variances are stable across time. Component loadings close to zero predominantly stay near to zero, and those that are notably further away do not converge to zero.

Figure 3.42: Place of Employment - Component 1: 2009 to 2016

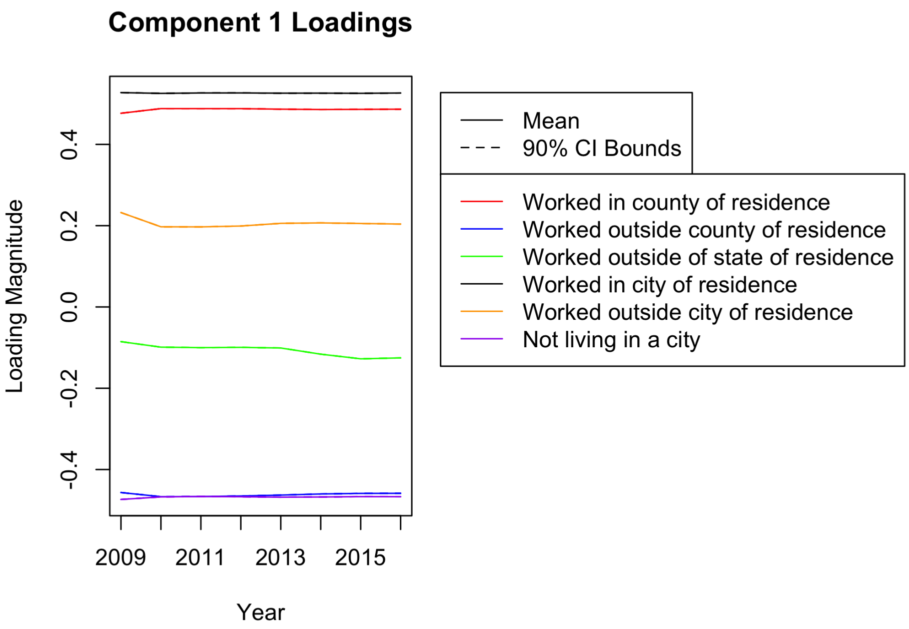


Figure 3.43: Place of Employment - Component 2: 2009 to 2016

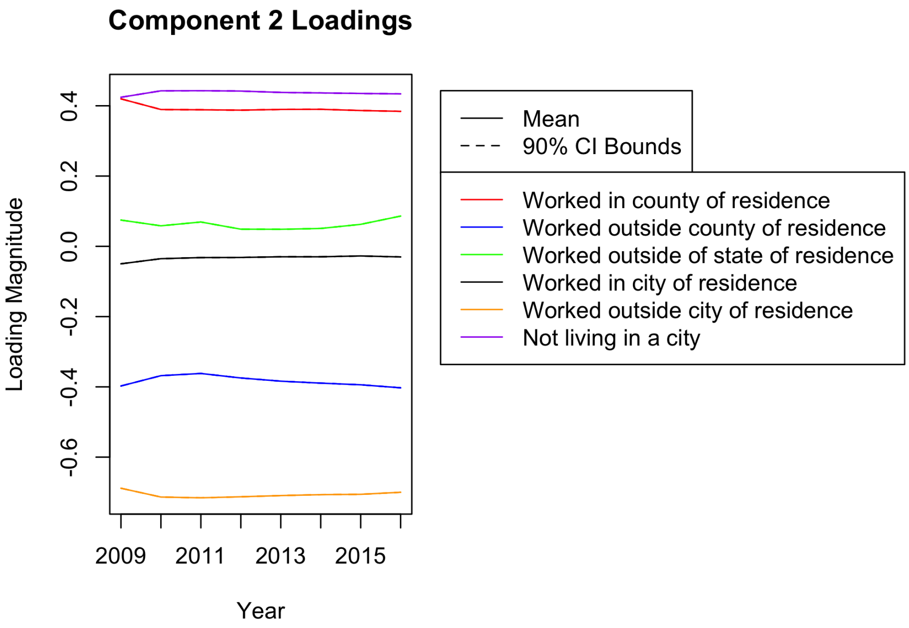


Figure 3.44: Place of Employment - Component 3: 2009 to 2016

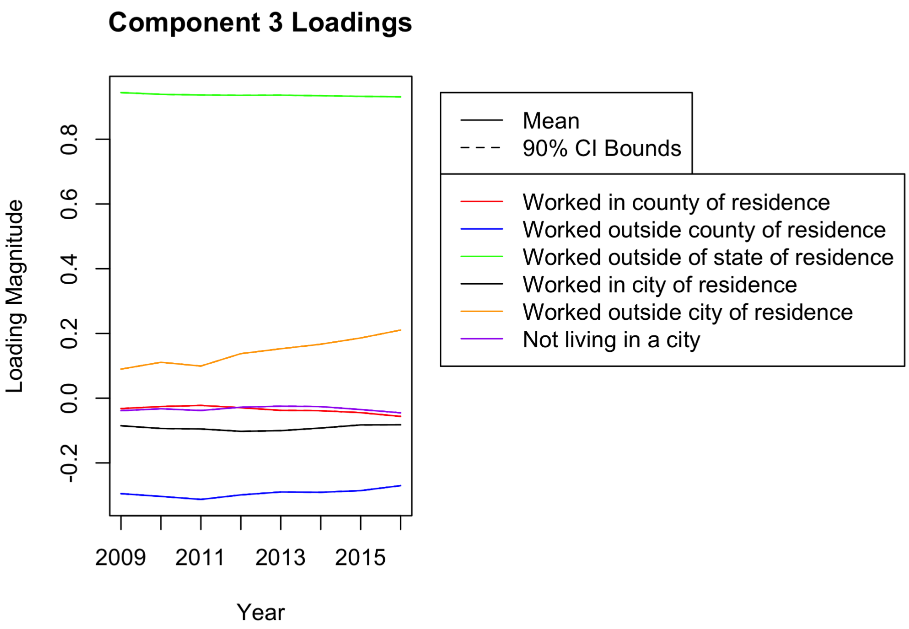
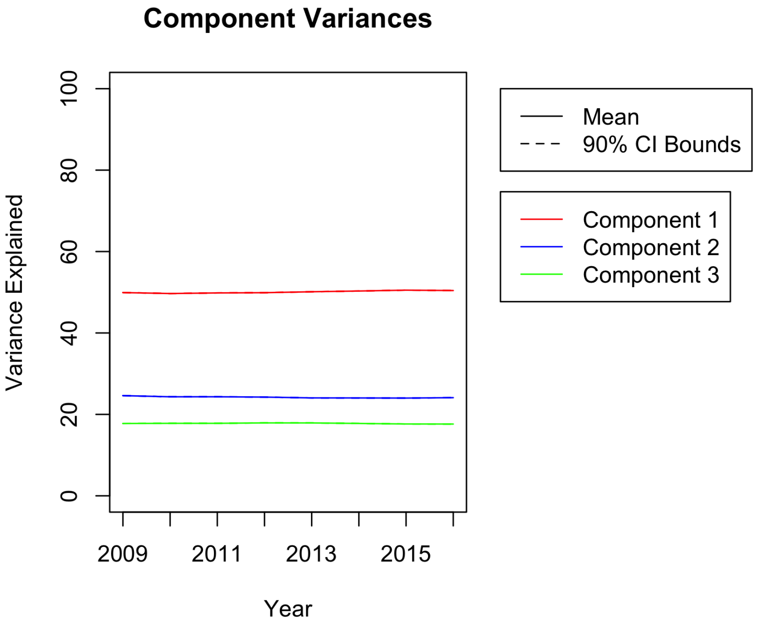


Figure 3.45: Place of Employment - Variance Explained: 2009 to 2016



3.6 Commuting Characteristics: Commute Times

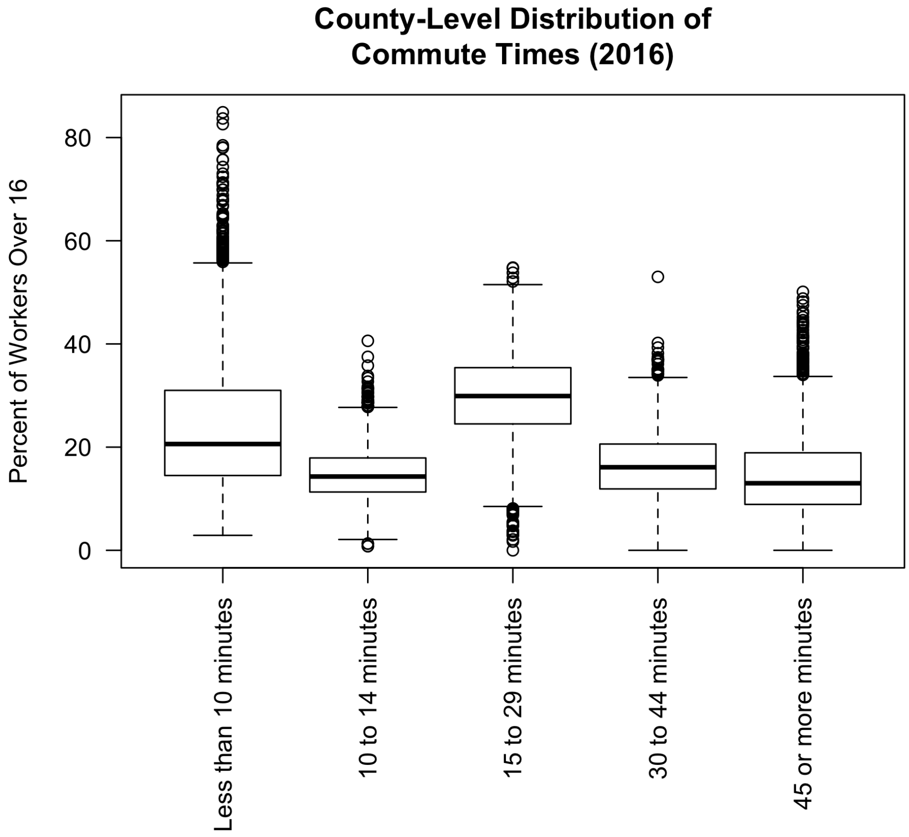
Commuting time is distinct from the geographical spread investigated in the previous section. The Census aggregates this data into nine time intervals. Preliminary principal component analysis shows that these can be condensed into five groups while maintaining the same patterns exhibited by the full data set.

Table 3.15: Commuting Time Intervals

Less than 10 minutes	10 to 14 minutes
15 to 29 minutes	30 to 44 minutes
45 or more minutes	

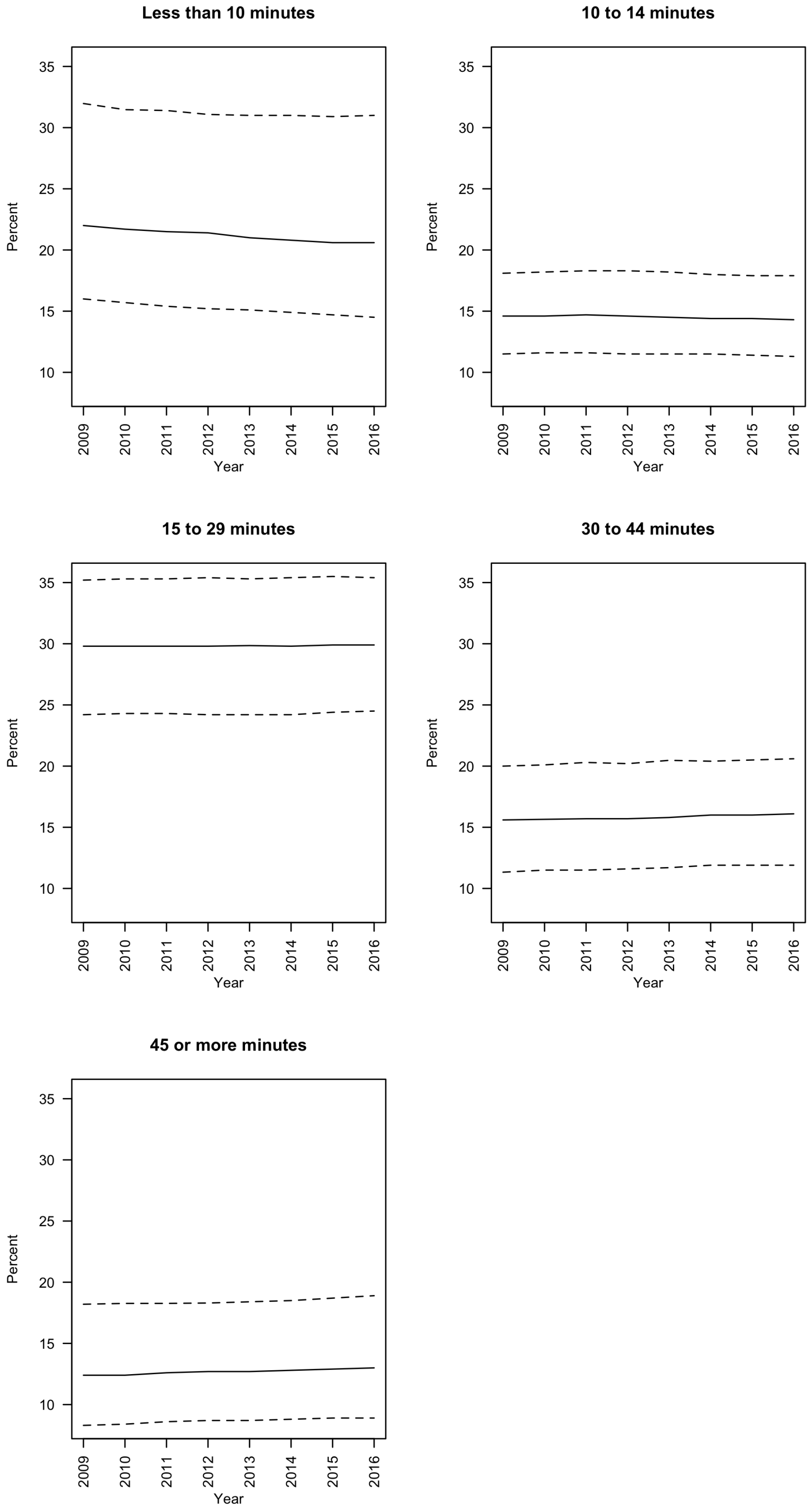
Commute times are moderately evenly distributed. No commute length has a county average of greater than 40%. People commuting between 15 and 30 minutes are in the most common group, followed by those who need less than 10 minutes. This shortest commute time group also has the largest spread, with some counties having most of their population in this category.

Figure 3.46: Commute Times Distribution across U.S. Counties



Examining these variables across time does not reveal strong trends, except possibly for commutes under 10 minutes. This is decreasing from 2009 to 2016, but this overall decline is also accompanied by a very wide interquartile range.

Figure 3.47: Inter-Quartile Ranges of Commute Times



3.6.1 Principal Component Analysis

The first component found in this data is for longer commutes in the positive direction, versus shorter in the negative. This distinction explains just over 50% of the variance present. The second strongest trend is for counties with commuters at either end of the spectrum, versus those with more people in the middle categories. This explains an additional 30% of all variation in the data set.

Table 3.16: Commute Times: Principal Component Loadings (2016)

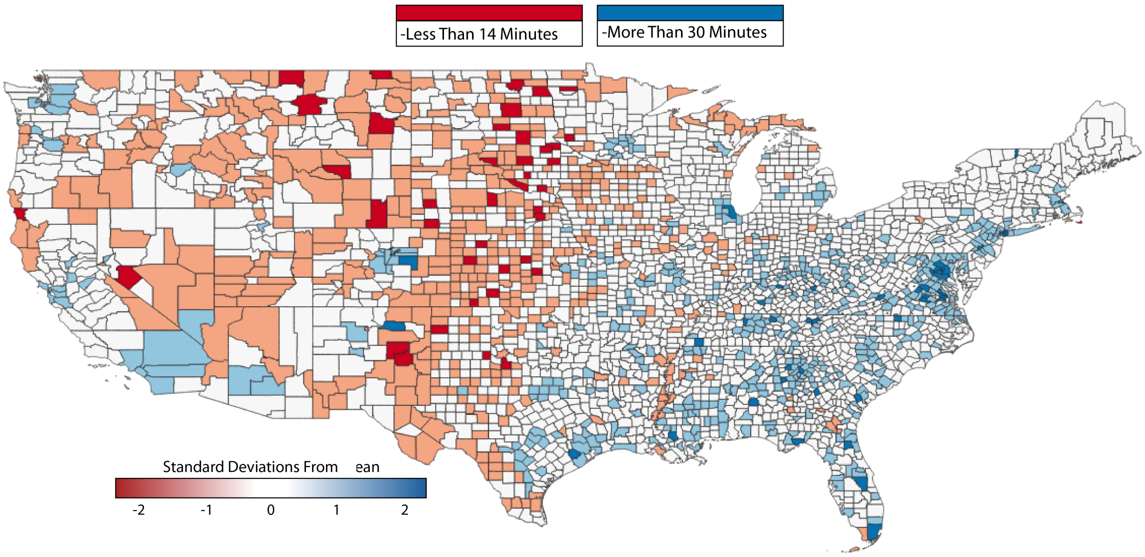
	Comp.1	**Comp.2**
Less than 10 minutes	-0.522	-0.394
10 to 14 minutes	-0.397	0.458
15 to 29 minutes		0.737
30 to 44 minutes	0.560	
45 or more minutes	0.479	-0.302

Table 3.17: Commute Times: Principal Component Variances (2016)

	Comp.1	**Comp.2**
Standard deviation	1.5943	1.2356
Proportion of Variance	0.509	0.305
Cumulative Proportion	0.509	0.814

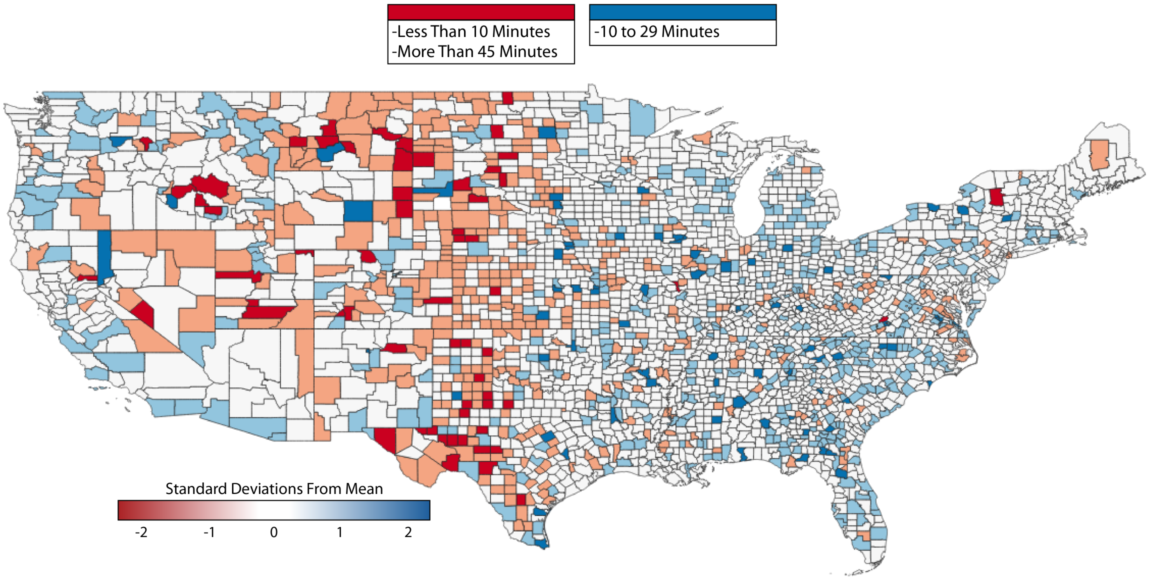
The first component shows a strong regional division. Much of the West and the western Midwest is dominated by commute times under 15 minutes. This is striking, and likely due the the low levels of concentrated urbanization throughout the area. This probability is reinforced by presence of counties with positive scores around major cities, for example Denver, Santa Fe, Seattle, and the San Francisco Bay Area. Going east, most counties are either associated with commutes over 30 minutes, or do not demonstrate a strong tendency in either direction.

Figure 3.48: Commute Times: Component 1 (2016)



The second component has a fairly similar geographic distribution to the first. The eastern half of the country is more strongly associated with commutes ranging from 10 to 30 minutes, and the western half has more counties with a bimodal distribution. Again, this could be caused by a more rural disposition, where larger distances between cities would push commutes towards being quite short, or quite long. Primary economic sector behavior is also strong in this portion of the country, and long commutes to reach locations for agriculture, mining, or oil extraction could be a driver behind this component. The pockets of strong positive scores (shown in blue) are often mid-size cities, such as Topeka, KS, Reno, NV, Montgomery, AL, and Greensboro, NC.

Figure 3.49: Commute Times: Component 2 (2016)



These results are robust with respect to time and resampling, remaining almost completely unchanged.

Figure 3.50: Commute Times - Component 1: 2009 to 2016

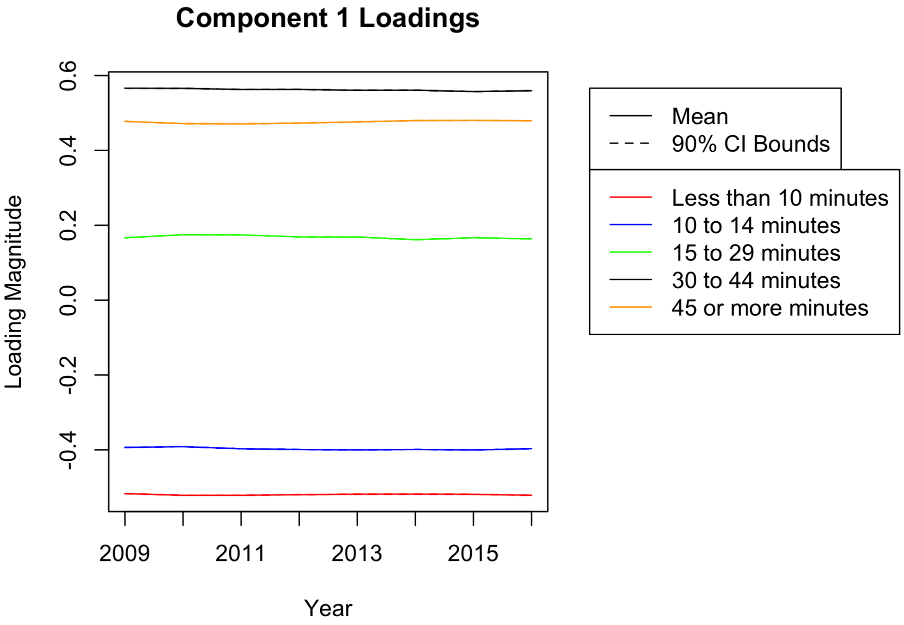


Figure 3.51: Commute Times - Component 2: 2009 to 2016

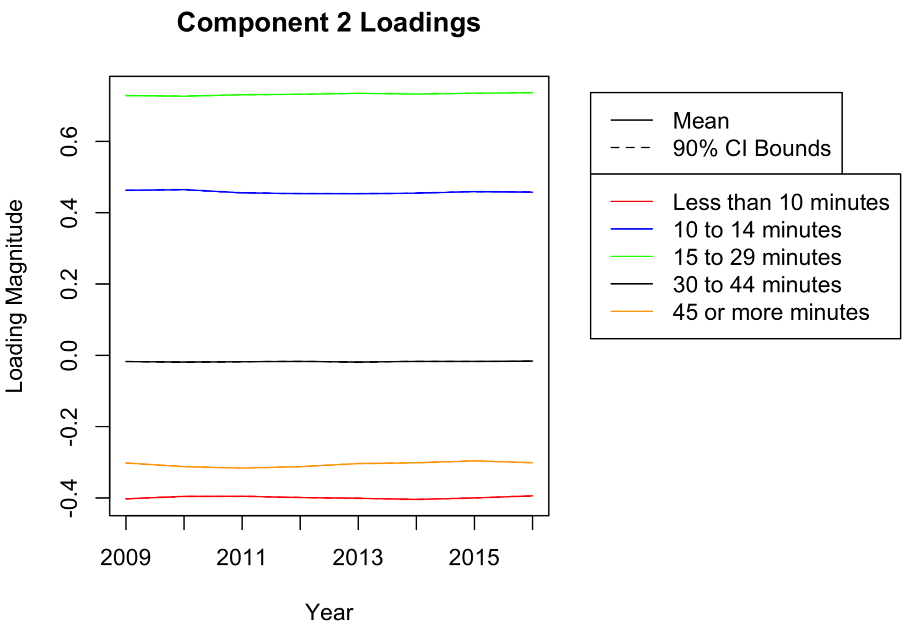
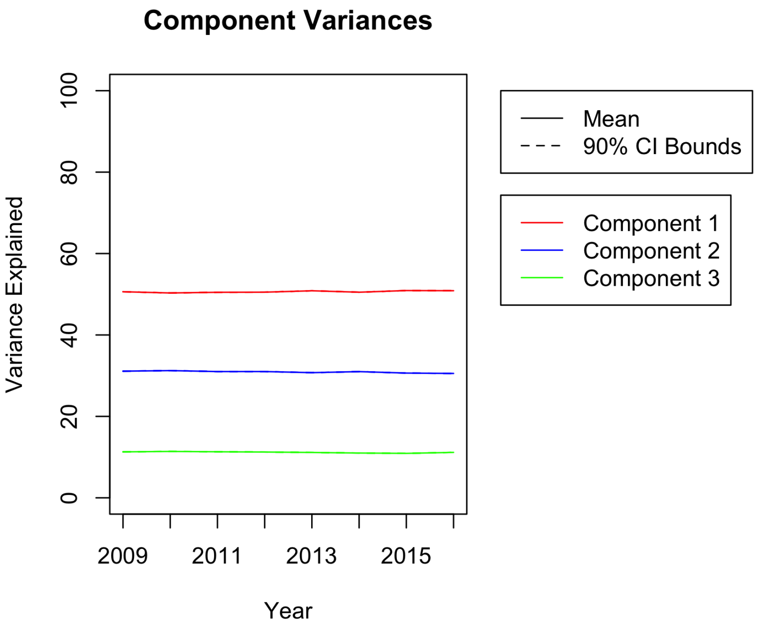


Figure 3.52: Commute Times - Component Variances: 2009 to 2016



3.7 Weeks Worked Annually

The number of weeks worked per year is a major part of of a persons job and potential income security. Principal component analysis finds four major groupings across this variable, as shown below.

Table 3.18: Weeks Worked Annually

Did not work	1 to 39 weeks
40 to 49 weeks	50 to 52 weeks

These data are for people aged 16 to 64, and on average amongst this population, employment is the norm. Around 55% of people are employed year-round, and under 30% of people on average are not working at all. There are, however, a number of outlying counties with high levels of people not employed, edging up even to a majority in some areas. These areas can be seen below, and include Appalachia, the lower Mississippi River region, as well as many rural counties across the country.

Figure 3.53: Percent of Population 16 to 64 Who Did Not Work (2016)

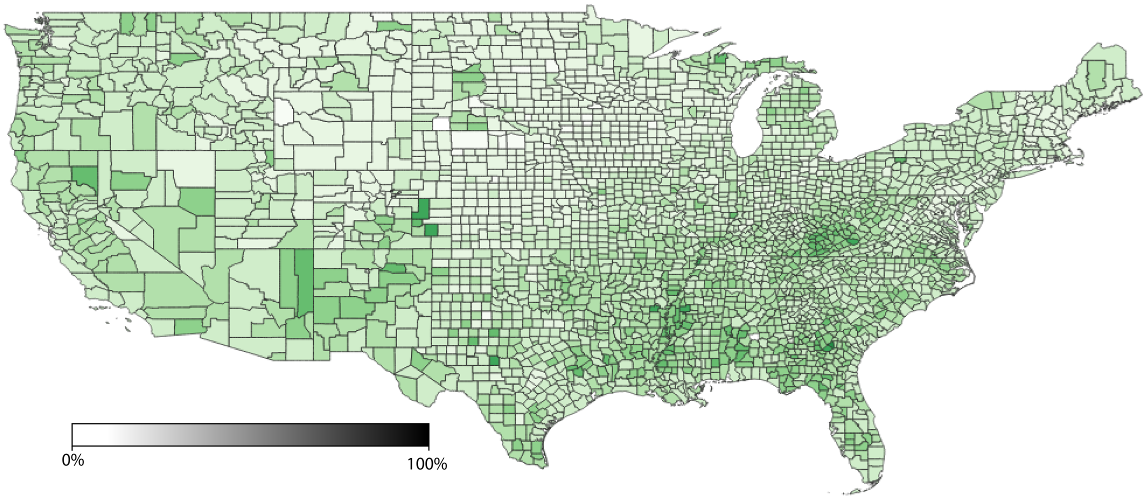
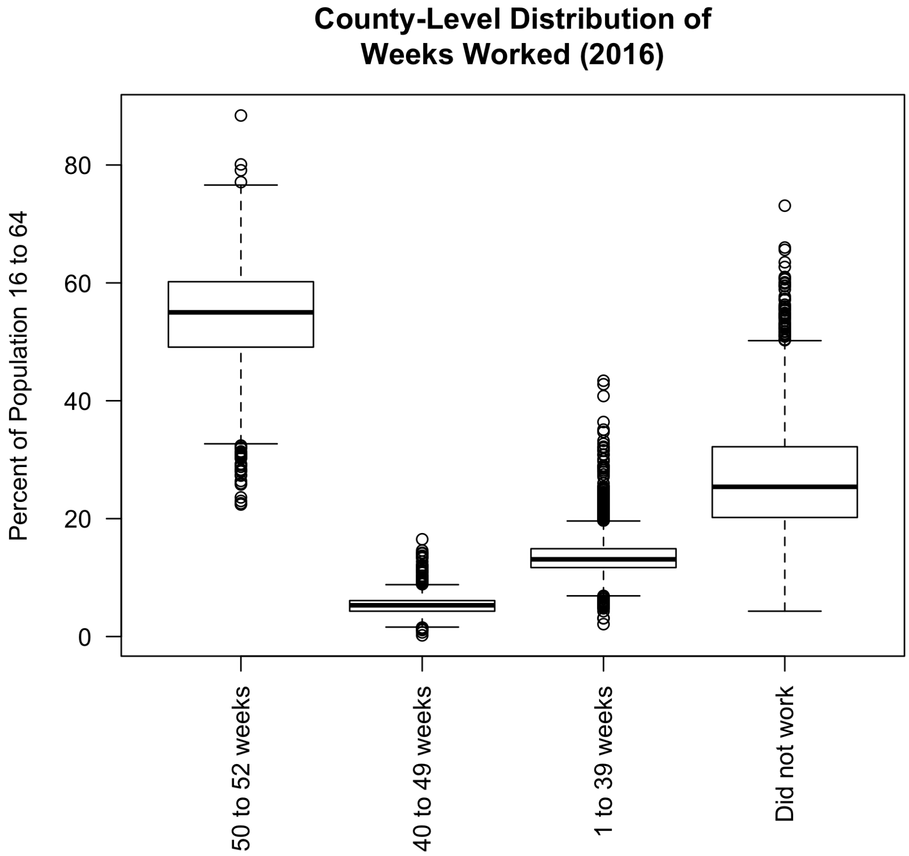
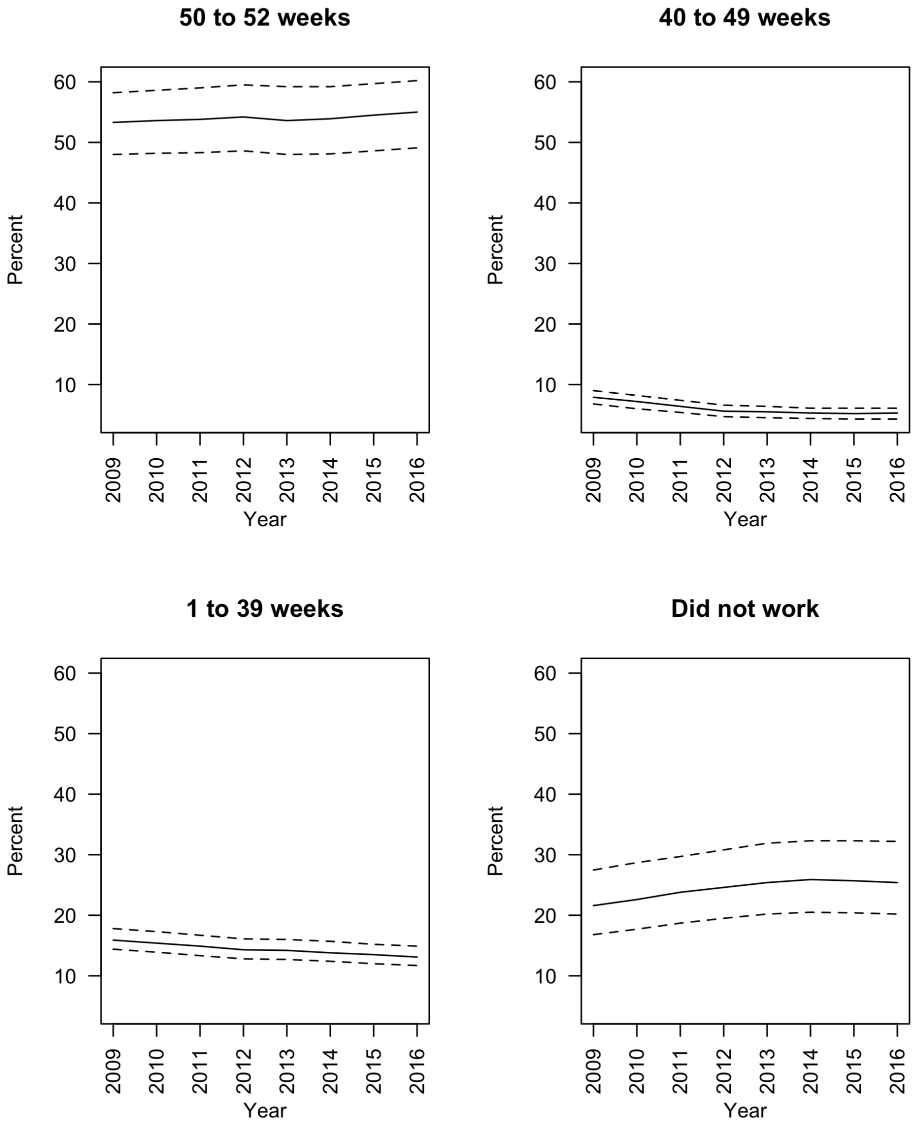


Figure 3.54: Weeks Worked Annualy (2016)



The percentage of people working year-round rose some over the study period, but the major observed trend here is an increase in people not working at all. This is deeply troubling, and could become a sustained pattern as long term unemployment can make reentry to the job market difficult [12, 13].

Figure 3.55: Inter-Quartile Ranges of Weeks Worked



3.7.1 Principal Component Analysis

The strongest trend found across counties in this data distinguishes between counties where more people work versus where more people do not. The second distinct underlying pattern contrasts counties where people work year-round, versus not. These account for 54.5% and 32.2% of variation respectively, for a total of 86.7% all told.

Table 3.19: Weeks Worked: Principal Component Loadings (2016)

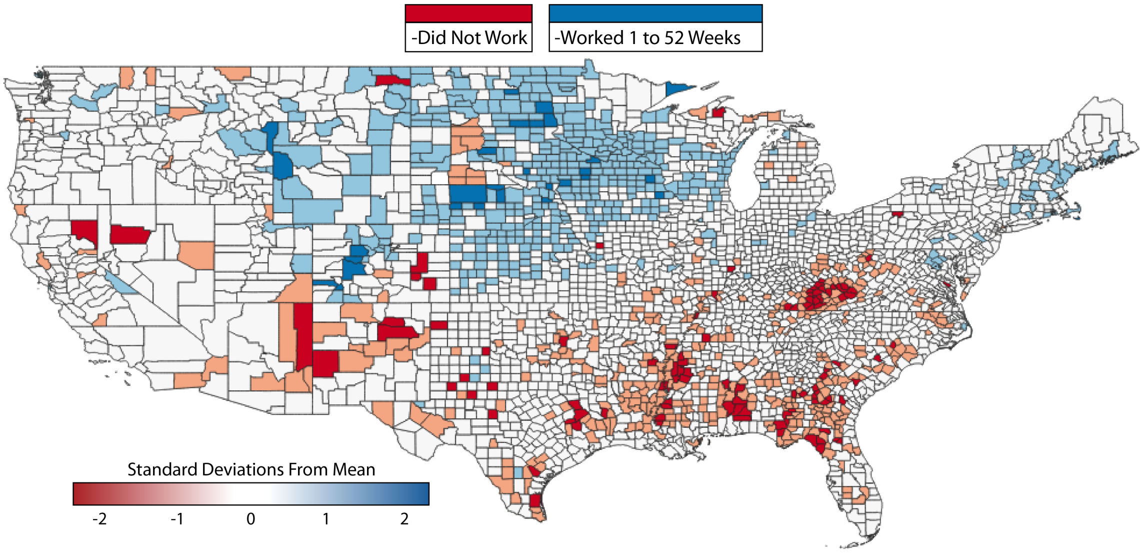
	Comp.1	**Comp.2**
50 to 52	0.534	0.531
40 to 49	0.455	-0.436
1 to 39	0.276	-0.707
Did not work	-0.657	-0.167

Table 3.20: Weeks Worked: Principal Component Variances (2016)

	Comp.1	**Comp.2**
Standard deviation	1.4767	1.1348
Proportion of Variance	0.545	0.322
Cumulative Proportion	0.545	0.867

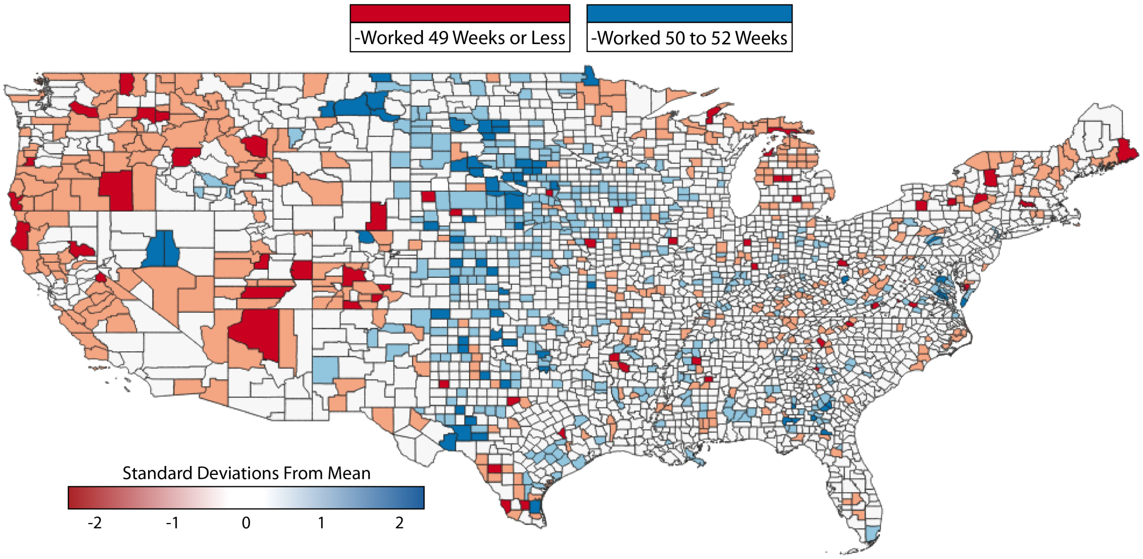
Much of the country is shown here in white. These counties do not differ greatly from the average distribution of weeks worked. However, a clear concentration of counties in blue can be seen in the upper Midwest, signifying higher proportions of people working on average. There are also hotspots of unemployment in the South and West, roughly matching those areas highlighted in the map earlier in this section.

Figure 3.56: Weeks Worked: Component 1 (2016)



Interestingly the second component shows that there is a clear line of counties in the Midwest with high proportions of people working all or almost all weeks of the year. As well, much of the West is characterized by less than year-round work, though referencing the previous map, this is likely not an indication of full unemployment.

Figure 3.57: Weeks Worked: Component 2 (2016)



The fluctuations observed in the first component over time serve only to emphasize that working between 1 and 39 weeks behaves more like the higher levels of employment as time goes on. For the second component, the low magnitude of the 'Did Not Work' group shows that the strongest contrast within this variable is between year-round employment and people with fewer weeks per year but who are still employed in some capacity.

Figure 3.58: Weeks Worked - Component 1: 2009 to 2016

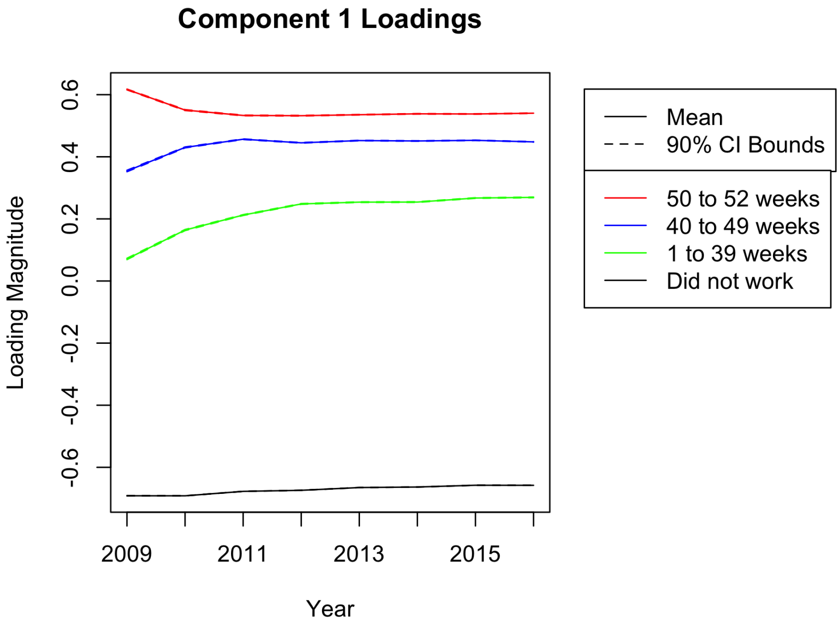


Figure 3.59: Weeks Worked - Component 2: 2009 to 2016

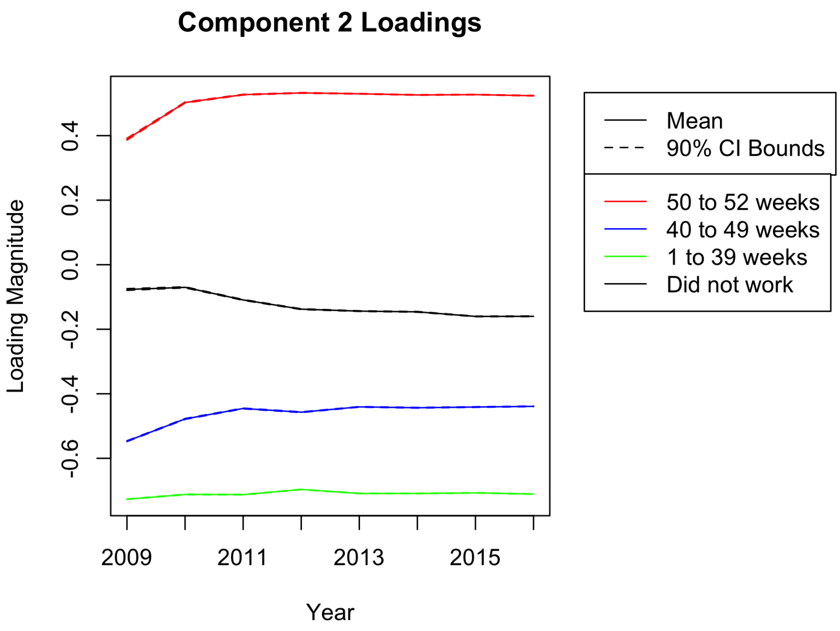
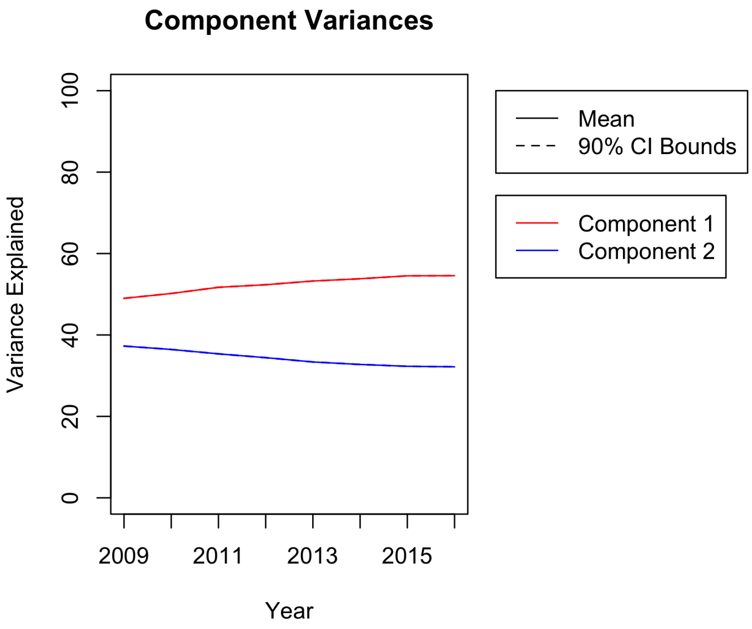


Figure 3.60: Weeks Worked - Proportion of Variance Explained: 2009 to 2016



3.8 Households and Families

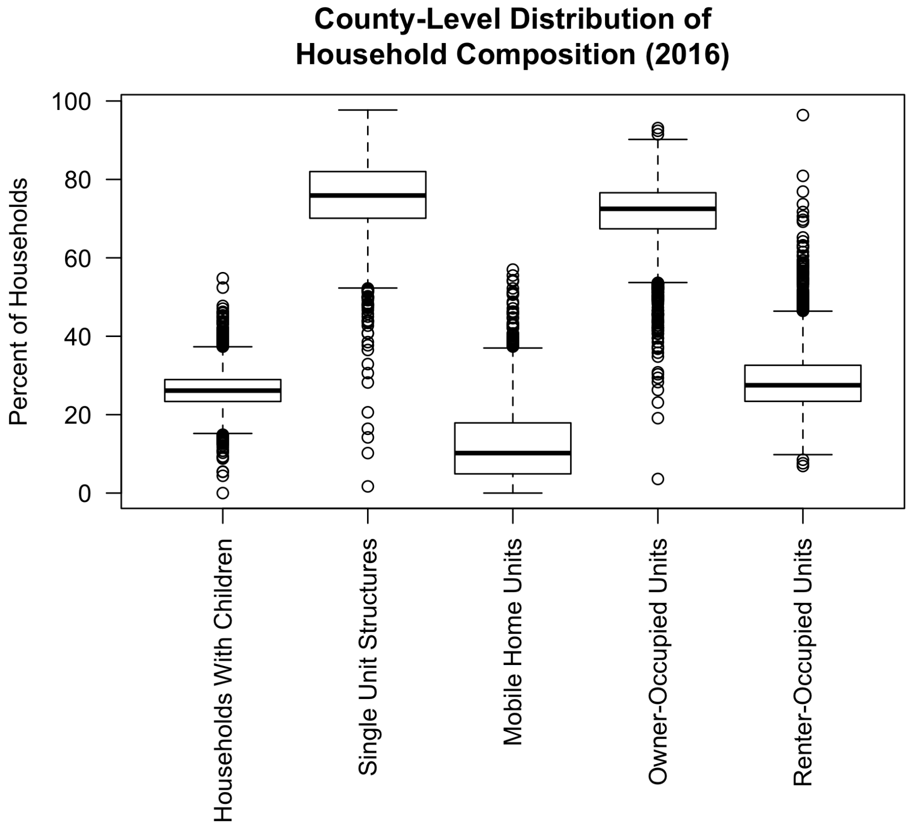
Household and family data give a sense of the social structure in a city or county. Here we look at the proportion of families that have children in a given county, the percent of households that are in single-unit structures (stand-alone houses), what percent of housing units are mobile homes, and the balance of renters versus owners.

Table 3.21: Household and Family Characteristics

Average Household Size	Households With Children
Single Unit Structures	Mobile Home Units
Owner-Occupied Units	Renter-Occupied Units

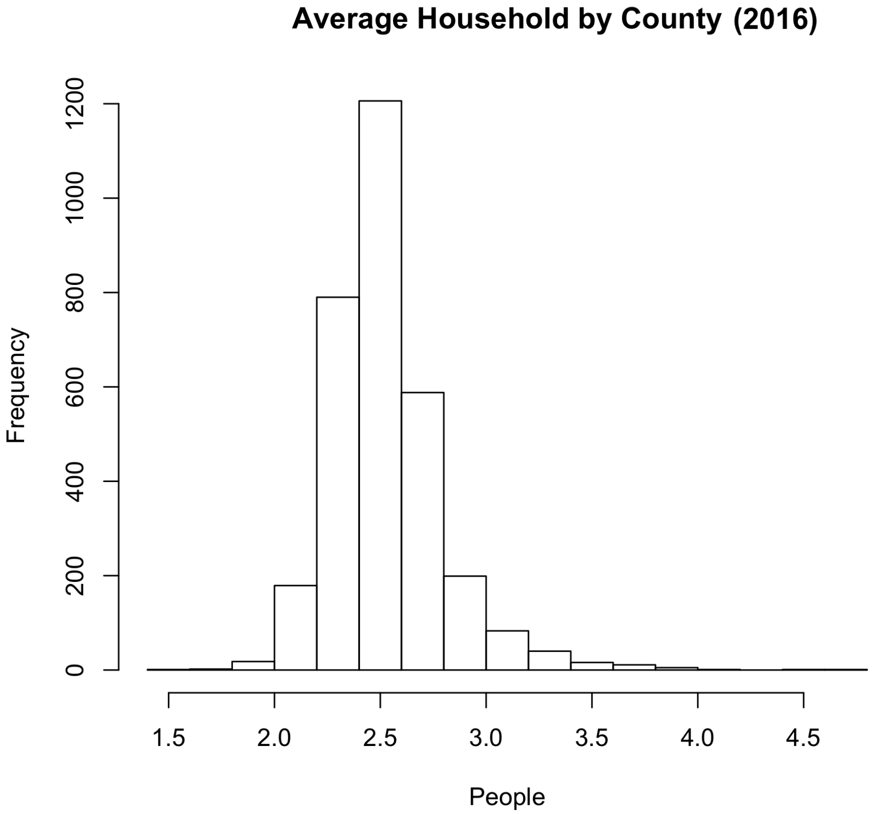
Interestingly, households with children generally make up a relatively small proportion of households, averaging about 30%. Single-unit structures on the other hand, are clearly the predominant housing mode, with an average of about 80% of all households. The long tail below the interquartile range for this variable hints at more densely populated counties, where more people are housed in multi-unit structures. Mobile homes are almost always minority for housing, but still average around 10% of all units. Finally, the balance between owners and renters is clearly skewed towards the owners, with overlap only in the outliers for each category.

Figure 3.61: Distribution of Households and Families (2016)



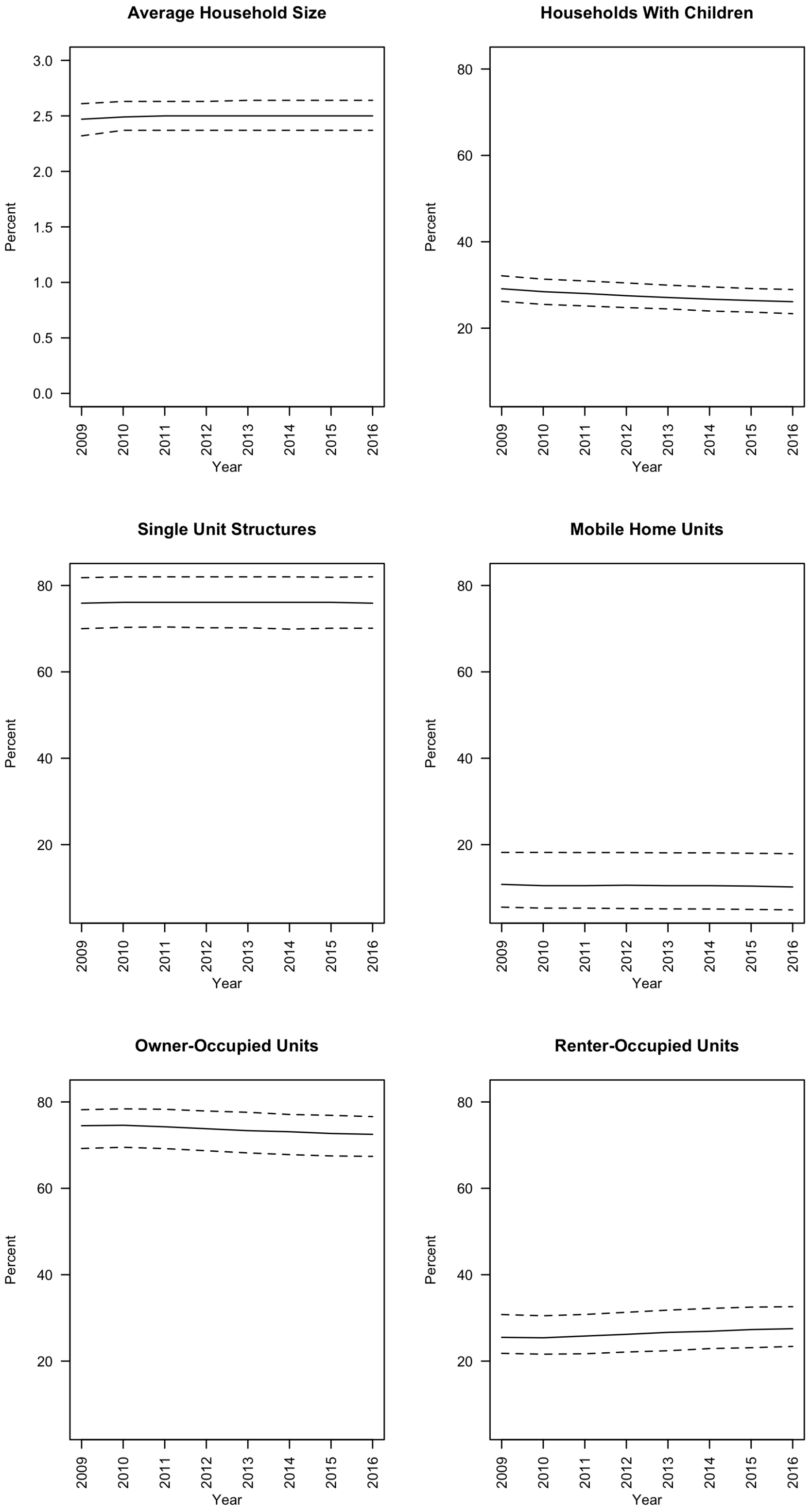
The distribution of county-level household size is centered on 2.5 people, and skews right. Not surprisingly, almost no county has an average house size below 2.

Figure 3.62: Average Household Size (2016)



The relative levels of single-units and mobile homes stay constant over the study period. On the other hand, the percent of households with children is dropping. There are also fewer owners, with a corresponding increase in renters, though from a low starting proportion of renters.

Figure 3.63: Inter-Quartile Ranges of Households and Families



Average Household Size

Households With Children

Single Unit Structures

Mobile Home Units

Owner-Occupied Units

Renter-Occupied Units

3.8.1 Principal Component Analysis

The strongest underlying pattern in this data is an axis between counties with more renting, large households with children, versus counties with a high proportion of single-unit houses that are owned by their occupants. Around 41% of variance is described by this distinction.

The second major axis found contrasts mobile homes with single-unit households that have more people and children. This trend explains just over a quarter of variation across all the variables.

The third component emphasizes the importance of mobile homes as a distinguishing trait because of its very high relative loading. The positive direction for this component finds mobile homes with large households, versus single-unit renting households in the negative direction. Together with the second component, this analysis shows that having more than the average percentage of mobile homes can alone be a major driver of principal component scores for a county.

Table 3.22: Families and Housing: Principal Component Loadings (2016)

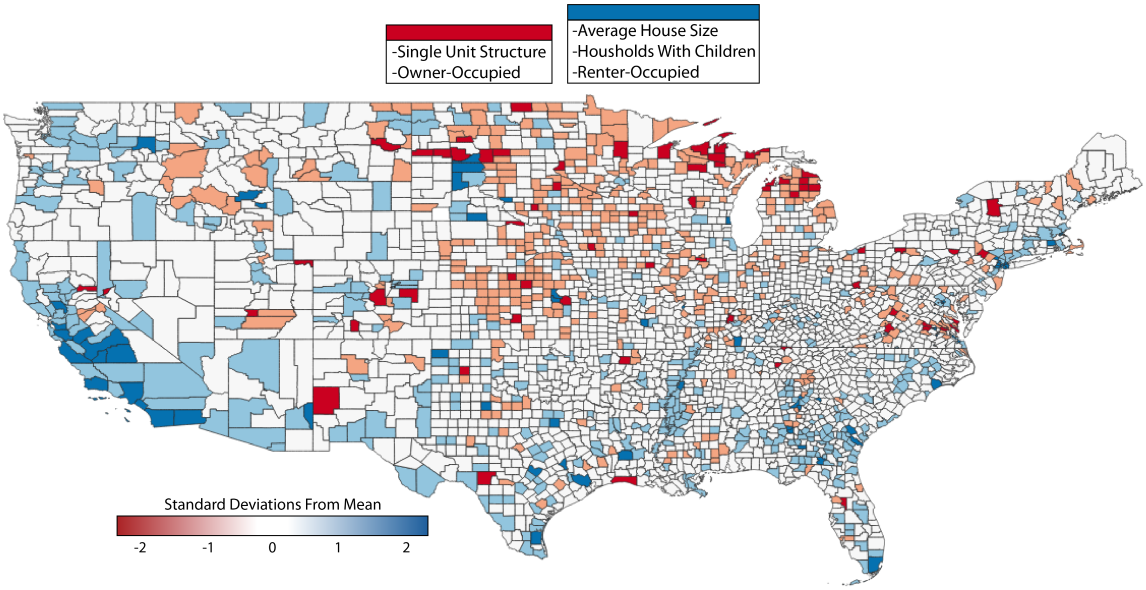
	Comp.1	**Comp.2**	**Comp.3**
Average Household Size	0.257	0.522	0.444
Households With Children	0.247	0.661	
Single Unit Structures	-0.420	0.376	-0.38
Mobile Home Units		-0.35	0.703
Owner-Occupied Units	-0.591		0.265
Renter-Occupied Units	0.59		-0.269

Table 3.23: Families and Housing: Principal Component Variances (2016)

	Comp.1	**Comp.2**	**Comp.3**
Standard deviation	1.5699	1.2563	1.1939
Proportion of Variance	0.411	0.263	0.238
Cumulative Proportion	0.411	0.674	0.912

The first map shows that the only major concentration of positive scores for Component 1 is in California. There are plenty of other areas exhibiting this trend, for example along the lower Mississippi river, but this component is much more dispersed than other trends. Counties with higher proportions of owner-occupied single-unit housing are mostly found in the northern and central parts of the midwest.

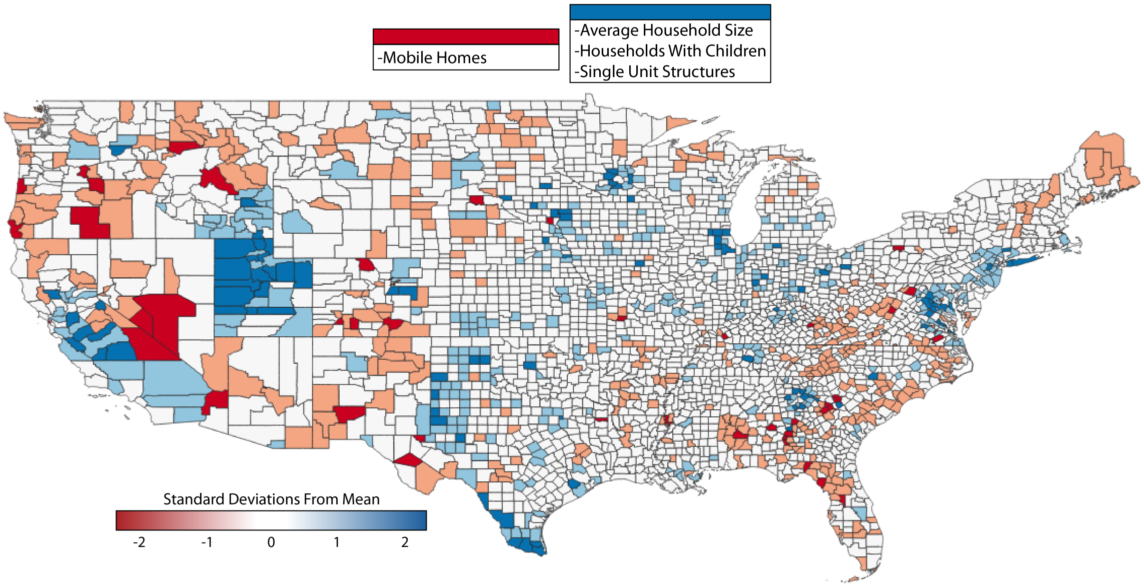
Figure 3.64: Families and Housing: Component 1 (2016)



Component 2 shows high numbers of mobile homes in red. These counties can be seen in some of the rural parts of Western states, though not all rural counties have positive scores here. There is also an area running from West Virginia to Florida that has a relatively high proportion these counties.

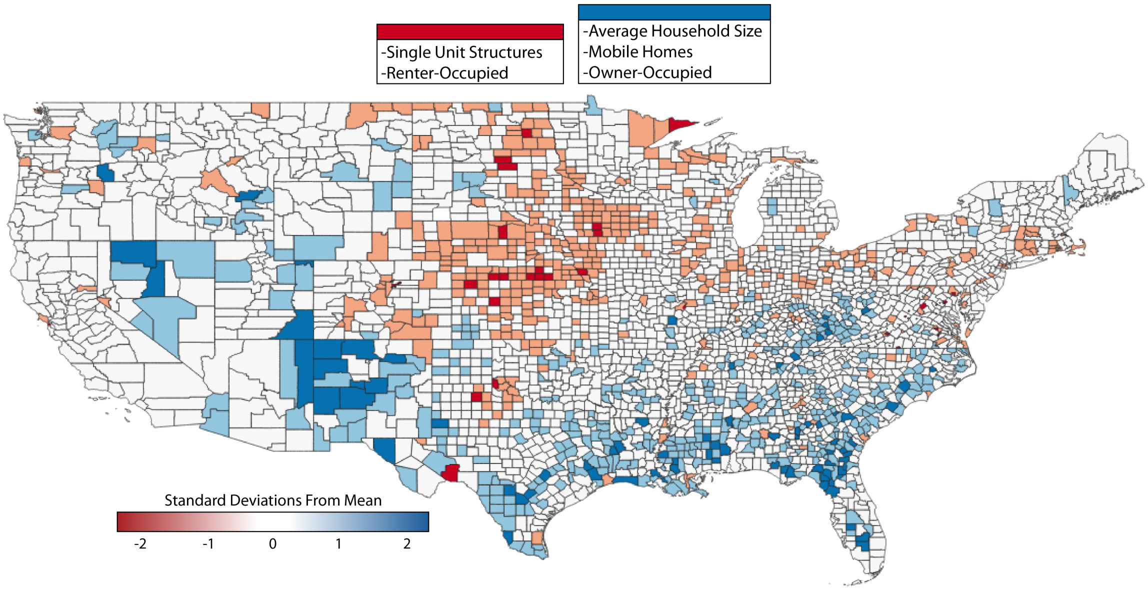
Blue counties on this map indicate more large families with children who live in stand-alone houses. Utah stands out the most clearly in this respect, but the suburbs of some major cities are also clearly present. For example, this can be seen around Chicago, Washington, D.C., and Minneapolis-Saint Paul.

Figure 3.65: Families and Housing: Component 2 (2016)



Negative scores for component three indicate an increased presence of people renting single-unit houses. These counties are overwhelmingly found in the upper and central Midwest. Positive scores find larger proportions of mobile homes and larger household sizes, and are seen all along the Gulf coast, as well as in Appalachia, New Mexico, and rural Nevada.

Figure 3.66: Families and Housing: Component 3 (2016)



The constituent components of these analyses are not stable over time. Positive scores for Component 1 are more defined solely by the presence of renters the further back one looks, de-emphasizing large households and the presence of children as important traits. For the second component, single-unit housing becomes a major driver of positive scores only by 2013-2014. The same is true for mobile homes with respect to negative scores. Component 3 relies less heavily on households with children, but picks up the effects of larger households more strongly as time goes on. Despite this shuffling of relative loading strengths, the variance explained by each component remains constant.

Figure 3.67: Families and Housing - Component 1: 2009 to 2016

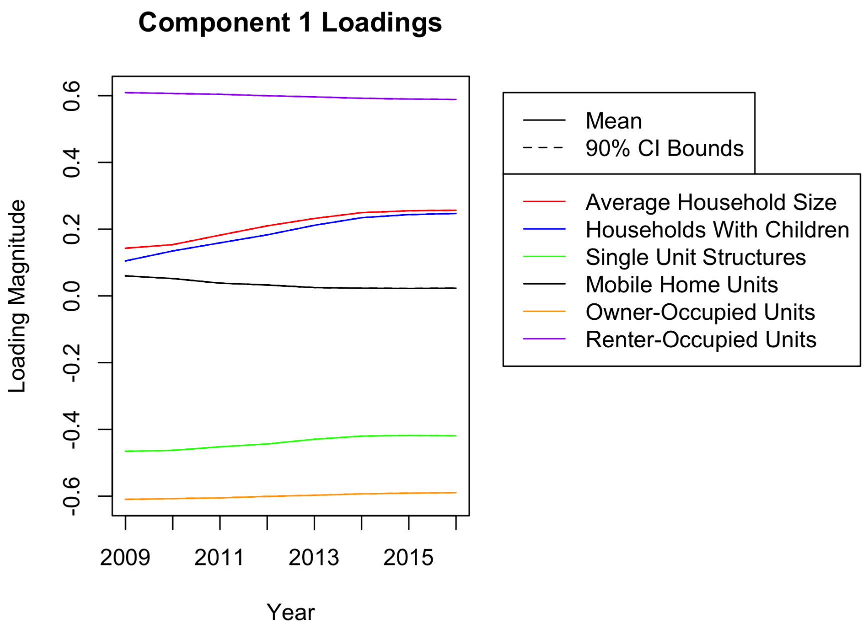


Figure 3.68: Families and Housing - Component 2: 2009 to 2016

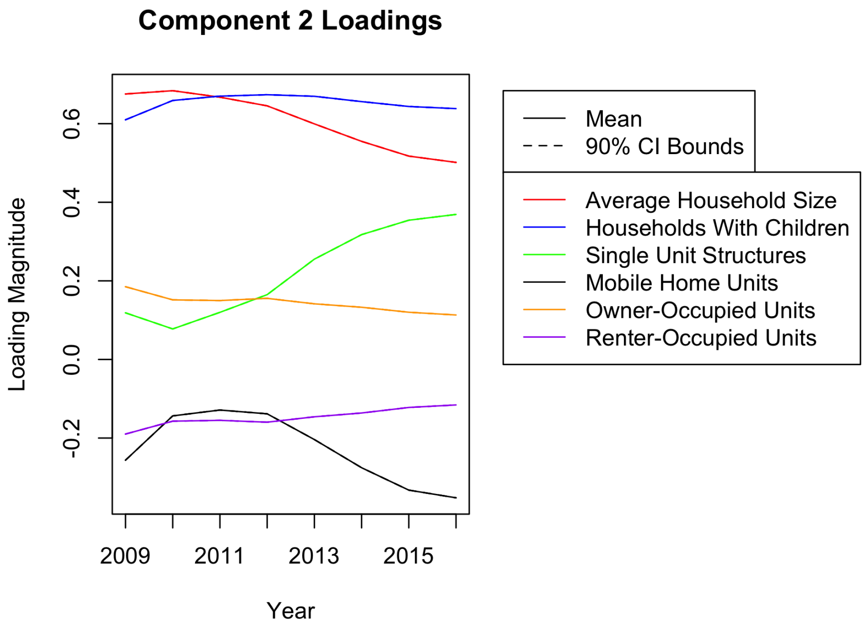


Figure 3.69: Families and Housing - Component 3: 2009 to 2016

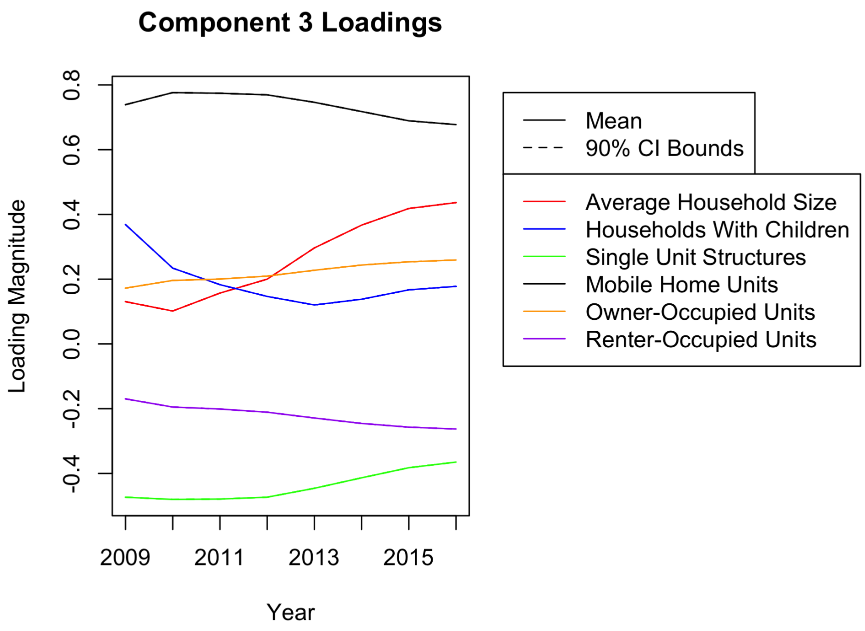
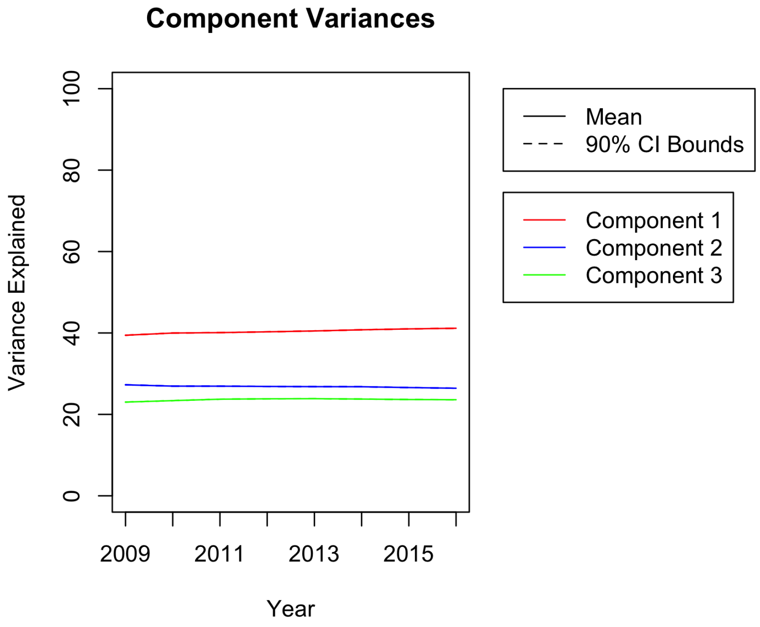


Figure 3.70: Families and Housing - Proportion of Variance Explained: 2009 to 2016



3.9 Marital Status

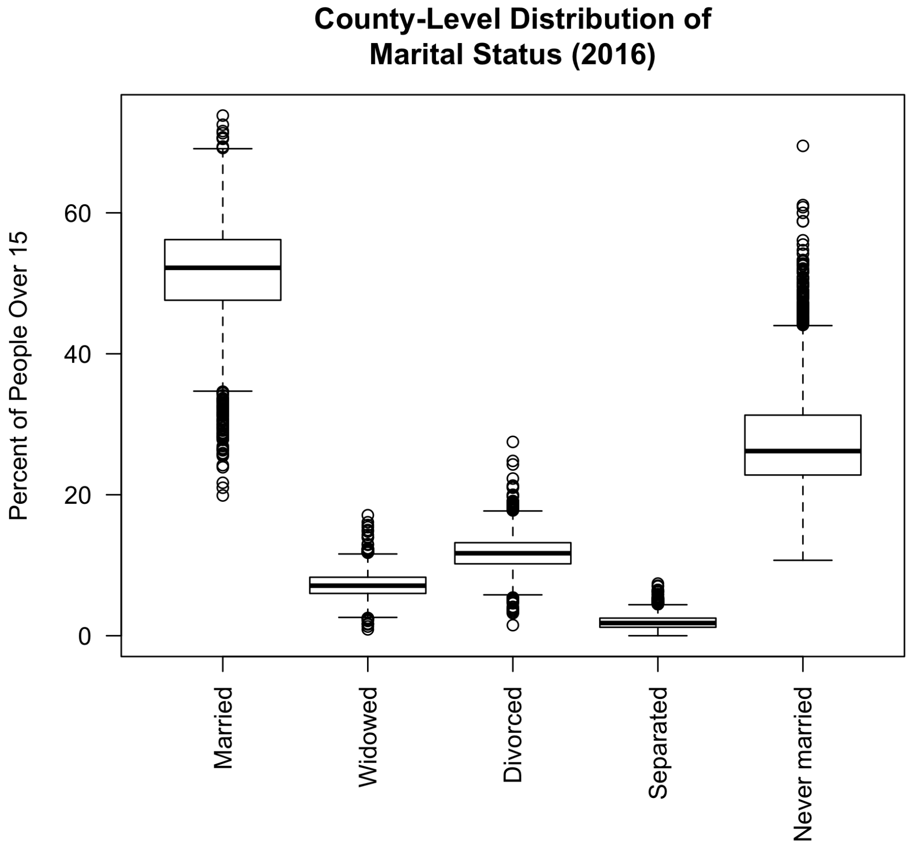
Five categories of marital status are recognized by the Census. This are exclusive categories, so people who are married but separated are in their own category, and are not counted in the 'Married' group.

Table 3.24: Marital Status

Married (Excluding Separated)	Never married
Divorced	Separated
Widowed	

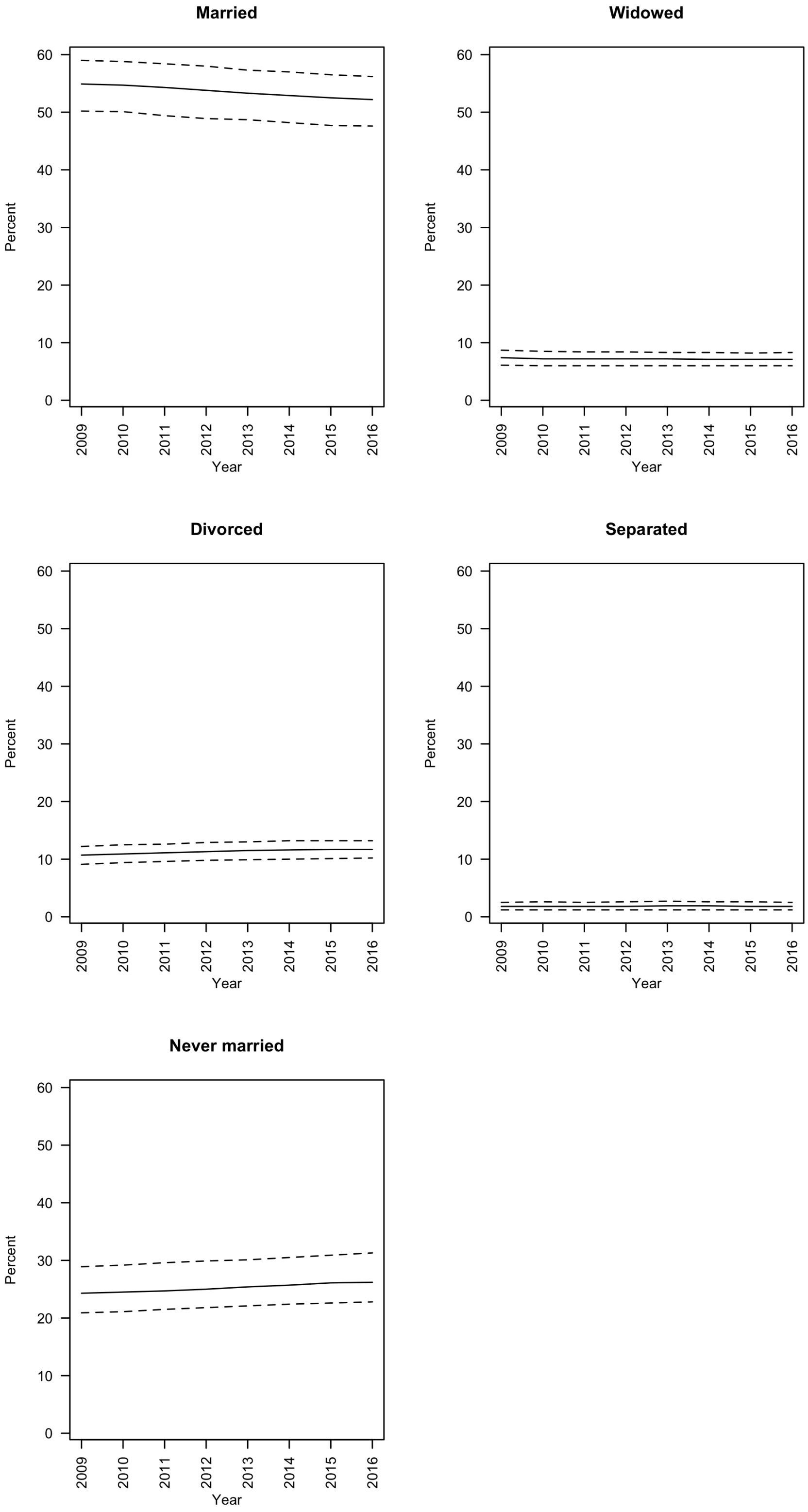
The most common marital state is that of marriage. On average, about half of the population by county is married. This is followed by people who have never been married, averaging around a third of the total population. The three other types of status never make up more than a fifth of a county population, with the exception of divorcees in a couple of counties.

Figure 3.71: Distribution of Marital Status (2016)



Over the period from 2009 to 2016, marriage rates were dropping, and never married rates are increasing. This hints at a trend of people postponing marriage, as opposed to getting divorced at higher rate, though divorce is edging up during this period as well.

Figure 3.72: Inter-Quartile Ranges of Marital Status



3.9.1 Principal Component Analysis

Predominantly, the distinction between counties is a tendency towards more married people, or more single, never married people. This accounts for 44% of the variance in the data. Interestingly, in this component, we see that people who are married but separated tend live in counties with more people who have never married. After this distinction is accounted for, the next strongest trend finds counties that have either more people who have never married, or more people who have been married but are now separated, divorced, or widowed. This accounts for almost 30% of all variation.

Table 3.25: Marital Status: Principal Component Loadings (2016)

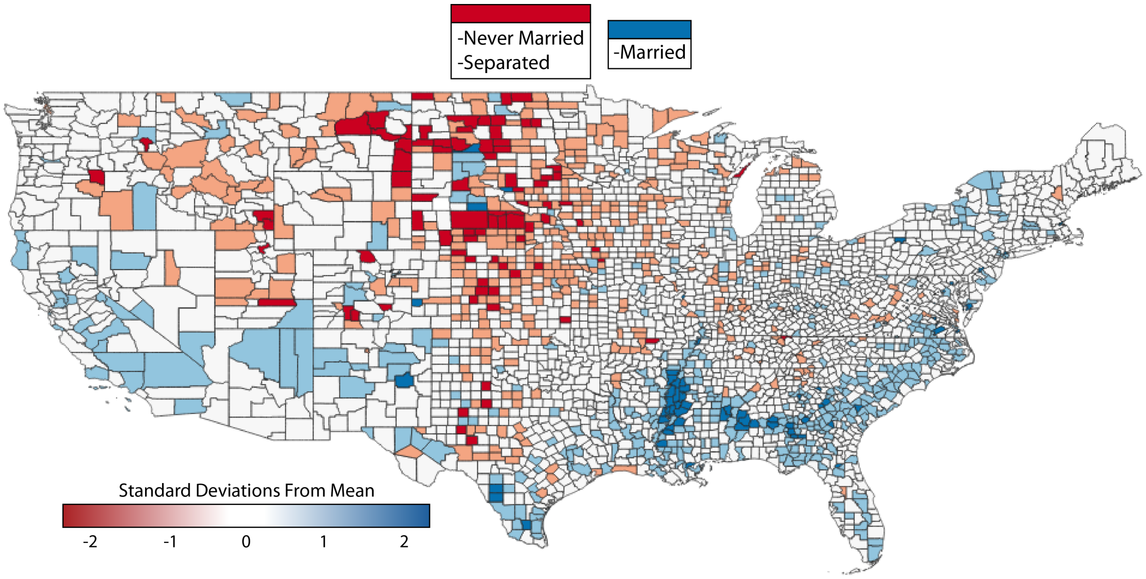
	Comp.1	**Comp.2**
Married	-0.646	
Widowed		-0.645
Divorced		-0.594
Separated	0.448	-0.325
Never Married	0.592	0.343

Table 3.26: Marital Status: Principal Component Variances (2016)

	Comp.1	**Comp.2**
Standard deviation	1.4845	1.2030
Proportion of Variance	0.441	0.290
Cumulative Proportion	0.441	0.730

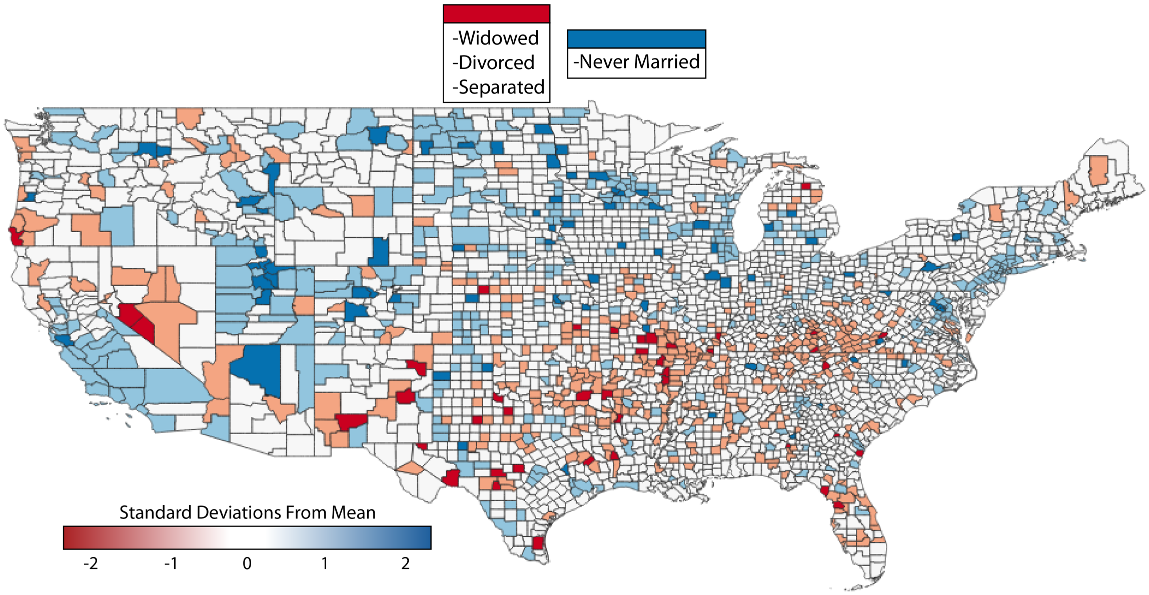
Mapping these patterns reveals a striking concentration of counties along the lower Mississippi river where there are more people who have never been married or who are separated, as well as sparser concentrations throughout the South. This contrasts sharply with the high rates of marriage in the northern and central Midwestern United States.

Figure 3.73: Marital Status: Component 1 (2016)



The second component is more dispersed geographically, but again the South is still the main area in one direction, in this case for people who are separated or no longer married. Large swathes of the West and northern Midwest are covered by counties where people are more likely not to have married yet. This trend can also be seen in places like Chicago, or the urban agglomeration stretching from Washington D.C. to New York City.

Figure 3.74: Marital Status: Component 2 (2016)



There is some movement in these component loadings across years, but no reversals or sign changes. The relative loadings of the two components are seemingly trending towards each other however, with the first getting less important over time.

Figure 3.75: Marital Status - Component 1: 2009 to 2016

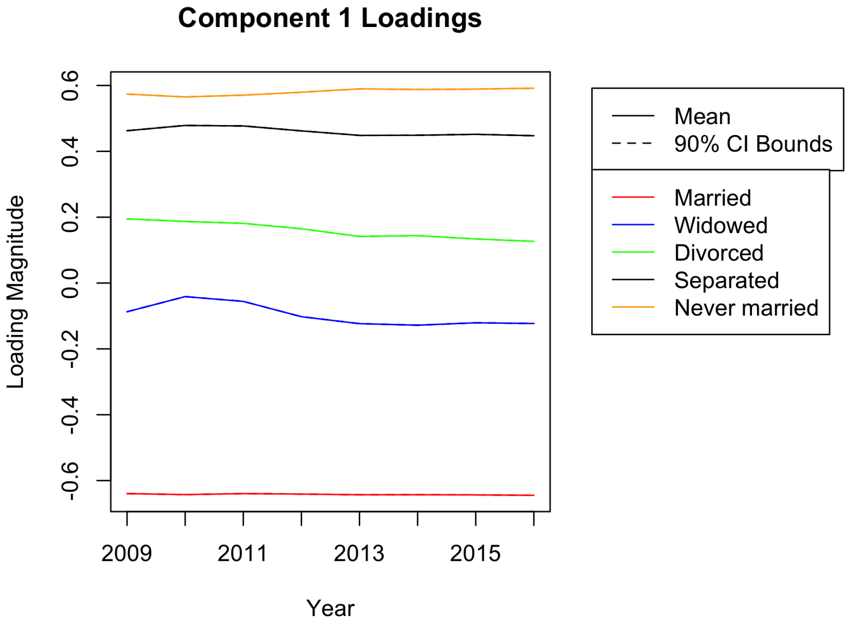


Figure 3.76: Marital Status - Component 2: 2009 to 2016

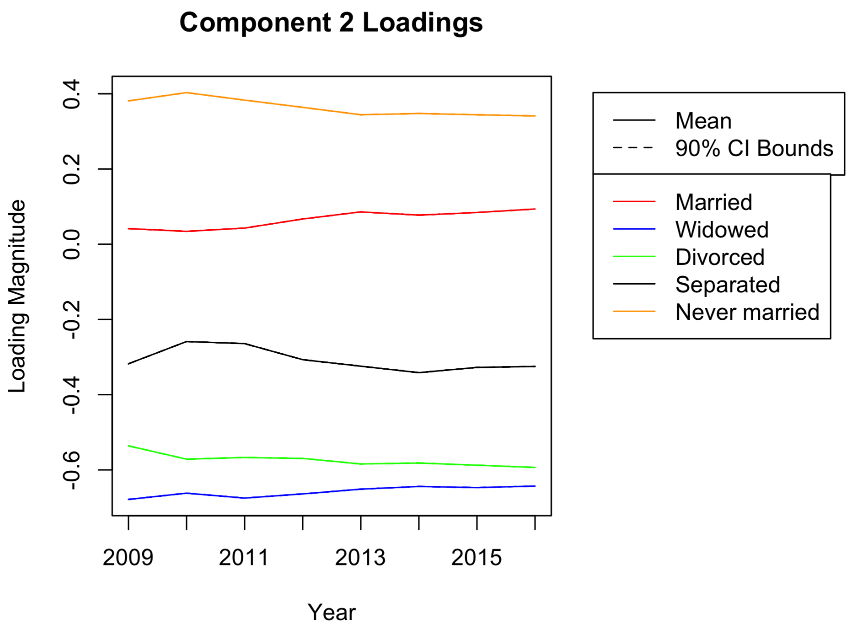
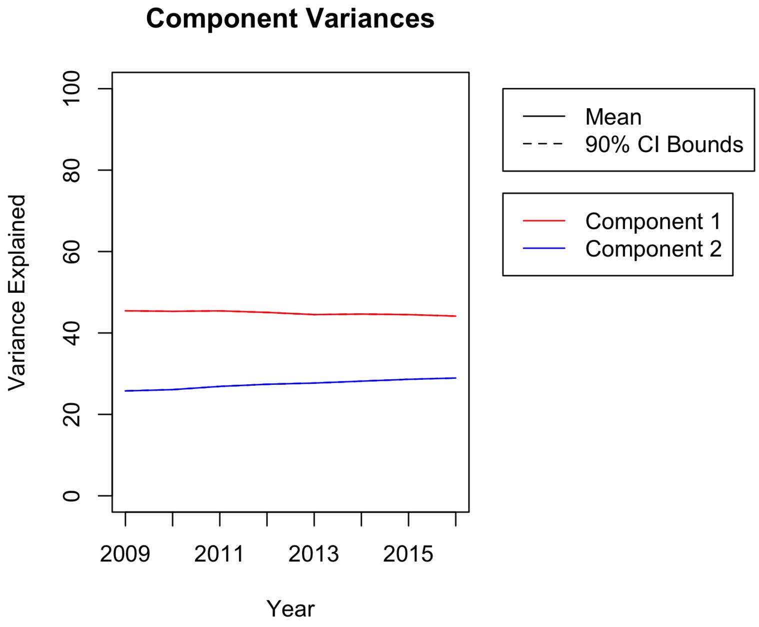


Figure 3.77: Marital Status - Proportion of Variance Explained: 2009 to 2016



3.10 Age

The Census groups age information into 17 5-year brackets, with an eighteenth bracket for people 85 years old and over. Exploratory principal component analysis shows that these 18 brackets can be well represented by the following age five groups. This exploratory analysis is detailed in the appendices.

Table 3.27: Age Groups

0 to 14	15 to 24
25 to 44	45 to 69
70 Plus	

The first plot below (Figure 3.78) shows the distribution of ages used in future analysis. However, caution must be used when examining this plot because not all of these groups cover equally sized age groups. A clearer distribution of the age distribution for the general population is shown in Figure 3.79.

Figure 3.78: Age Group Distribution

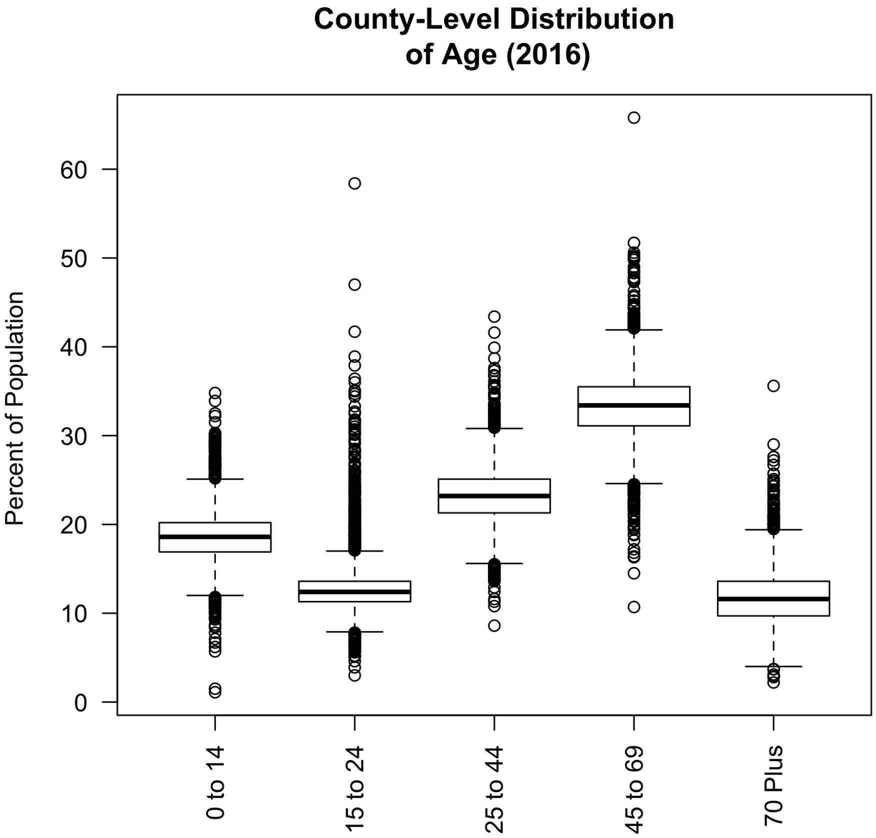
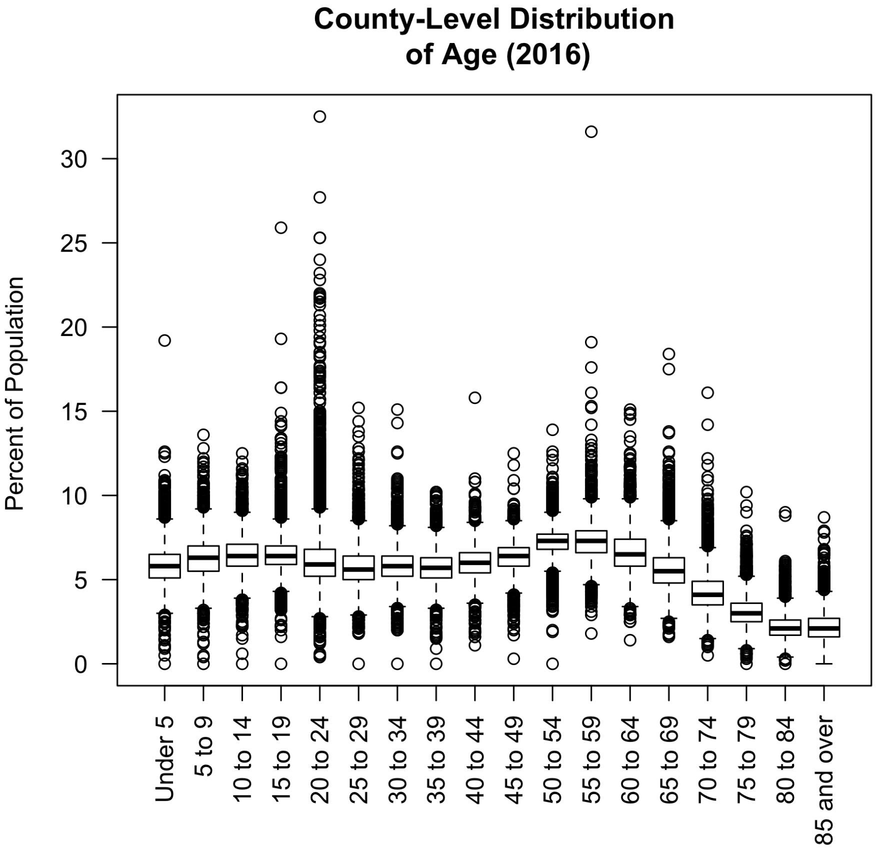


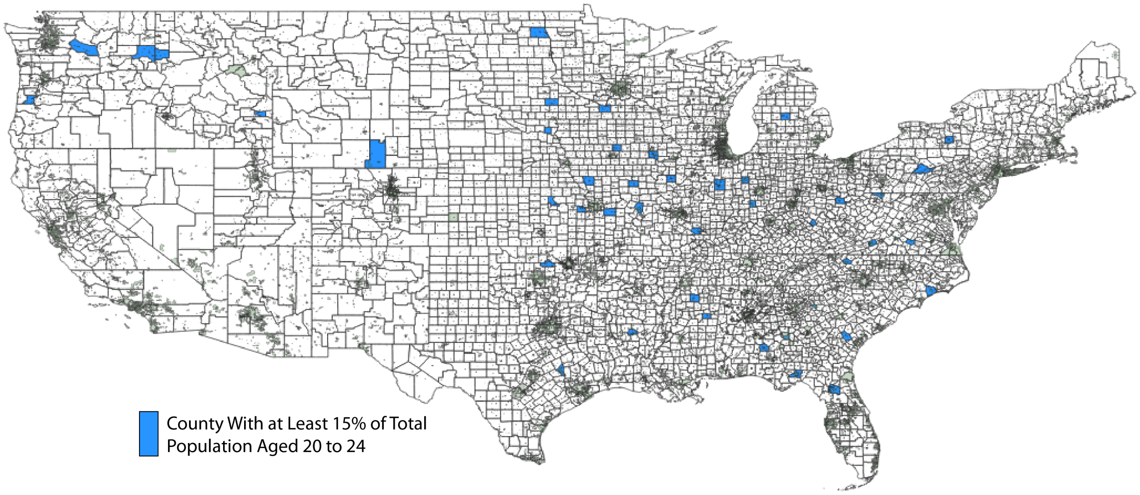
Figure 3.79: Full Age Group Distribution



Across the full array of age brackets, two peaks can be seen. The first, is for people aged 50 to 64. This age range is on average the most common, at least at the county level. The second peak in the distribution is for the outliers of younger adults, those aged 20 to 24. This is not reflected in the averages or inter-quartile ranges, but the strong presence of outliers in the higher percentages indicate that there are counties that disproportionately skew towards this specific five year age group.

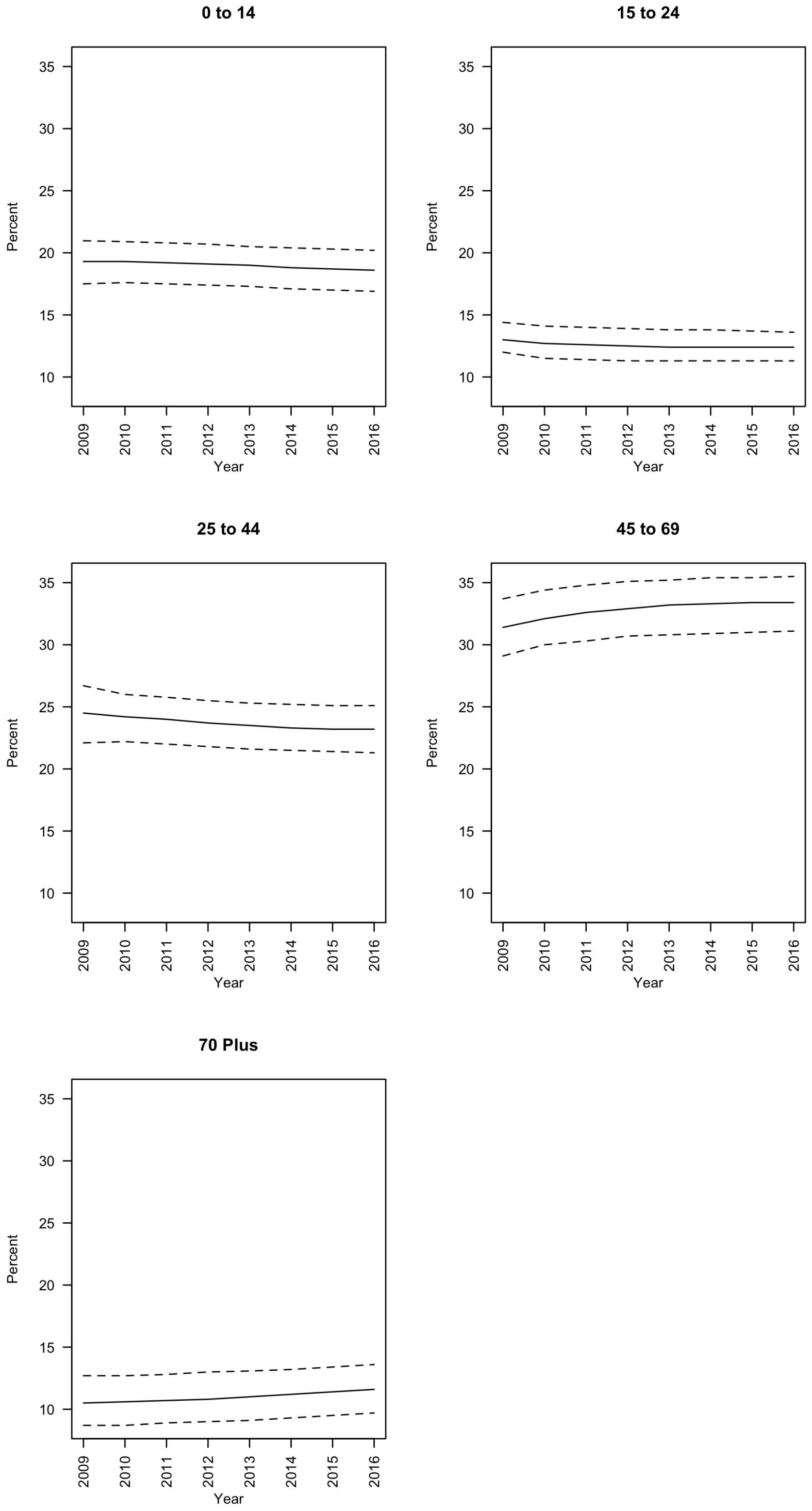
Interestingly, these counties are not in big cities, but instead are scattered throughout rural America. Shown below is a map of all counties for whom 15% or more of their population is aged 20 to 24.

Figure 3.80: Counties With at Least 15% of Population Aged 20 to 24



The clearest trends are in the top three age groups. The percentage of people between 25 and 44 is dropping slowly, while the proportion of people 45 to 69 initially ticked up seemingly more quickly than the former group decreased. This can only be due then to immigration or a shifting between counties. The average levels of people over 70 also are increasing over the study period, though 75th percentile of percentages is still lower than most other age groups.

Figure 3.81: Inter-Quartile Ranges of Age Groups



3.10.1 Principal Component Analysis

The dominating trend found in age group data shows counties 44 and under, against those 45 and over. This component alone accounts for more than 60% of all original variation. The next distinction finds counties tend to have either more children under the age of 15, or more people aged 15 to 24. Finally, a third axis is found distinguishing counties that tend more towards having people under 24 or over 70, versus having people in their working years, those aged 25 to 69.

Table 3.28: Age Groups: Principal Component Loadings (2016)

	Comp.1	**Comp.2**	**Comp.3**
0 to 14	-0.377	0.688	0.491
15 to 24	-0.410	-0.708	0.297
25 to 44	-0.432		-0.694
45 to 69	0.504		-0.299
70 Plus	0.499		0.317

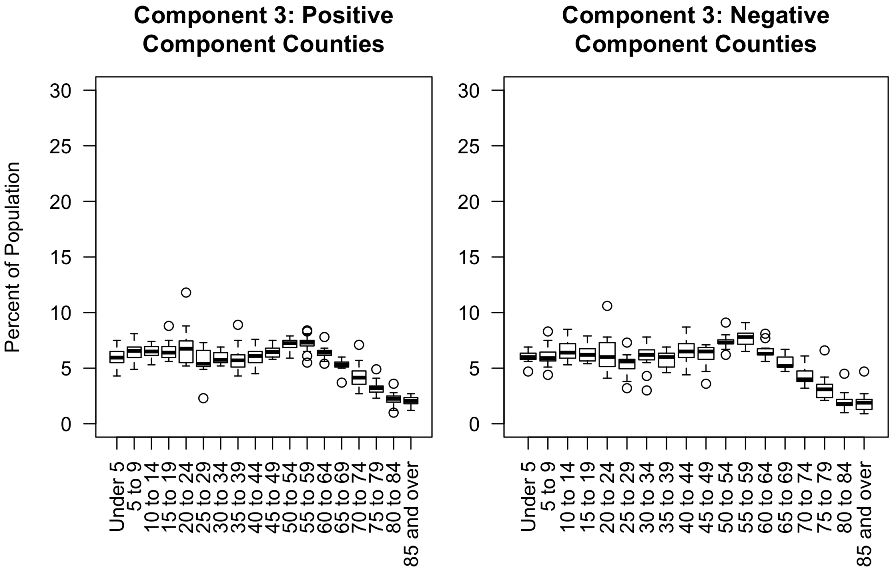
Table 3.29: Age Groups: Principal Component Variances (2016)

Comp.1	**Comp.2**	**Comp.3**	
Standard deviation	1.7676	0.8767	0.8587
Proportion of Variance	0.625	0.154	0.148
Cumulative Proportion	0.625	0.779	0.926

For the third component it is important to note that these trends are not exclusive, particularly as this component accounts for just under 15% of the total variation. In counties with positive scores, we would expect on average, more young people, but there will still be parents present for those young people. Similarly, in counties where more people between 25 and 69 are expected, the presence of children is not precluded. Below is a plot of the overall age distribution for counties that are at least one standard deviation in the positive direction away from zero for Component

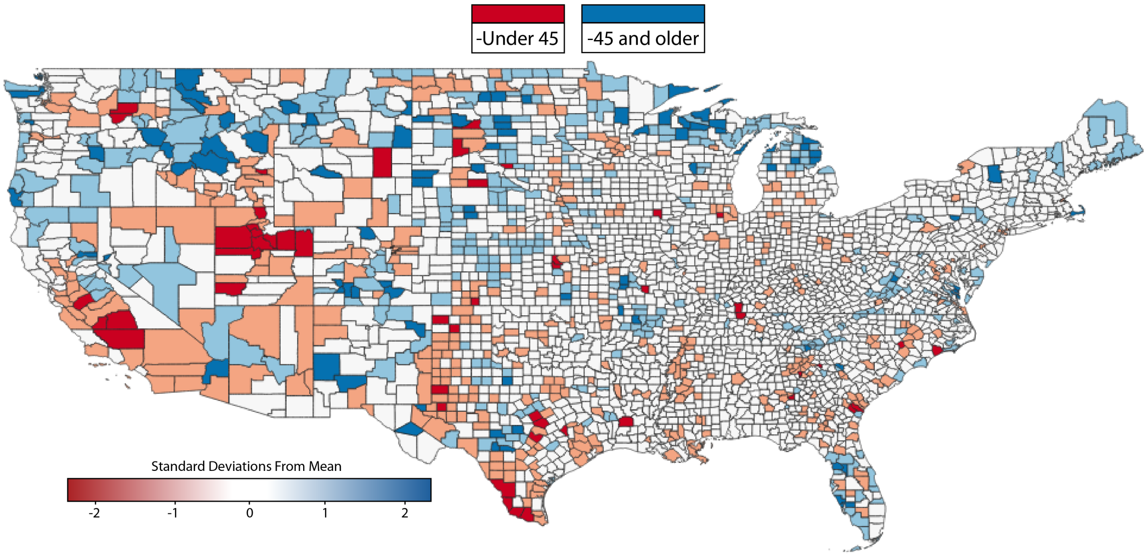
3, versus those that are at least one standard deviation in the negative direction. Clearly, the pattern identified in the third principal component is a subtle one.

Figure 3.82: Age Distribution Split by Scores for Component 3



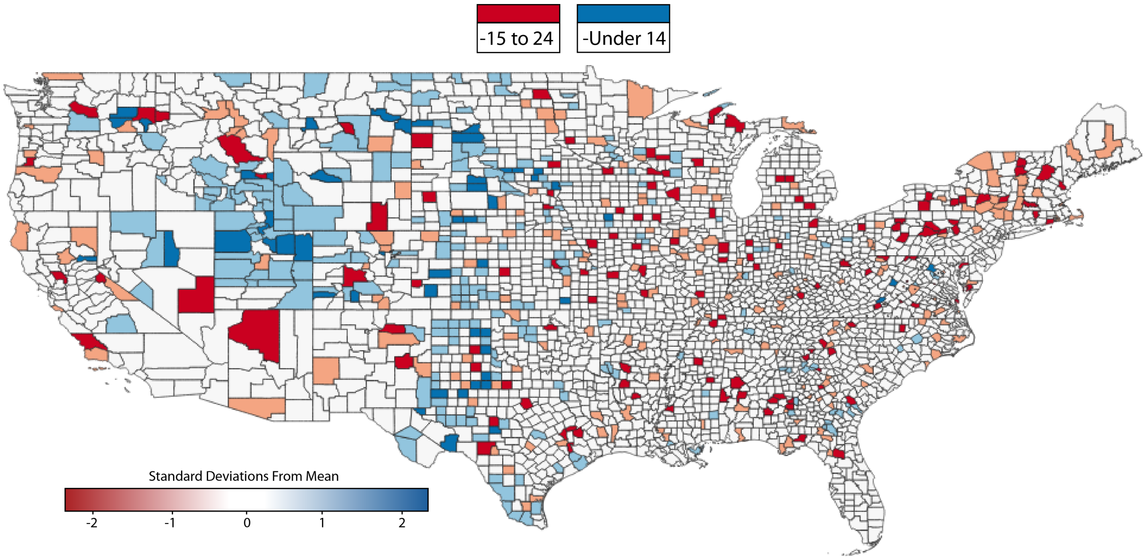
Interestingly, there appears to be a spread of counties across the northern United states where people are more likely than average to be over the age of 45, and less likely than average to be younger. People under 45 years old are more concentrated in Southern California, Utah, and parts of Texas.

Figure 3.83: Age: Component 1 (2016)



The second component shows a very striking pattern. Counties shown in deep red have much higher rates of people aged between 15 and 24 than average. This corresponds well to the earlier map depicting concentrations of this age group.

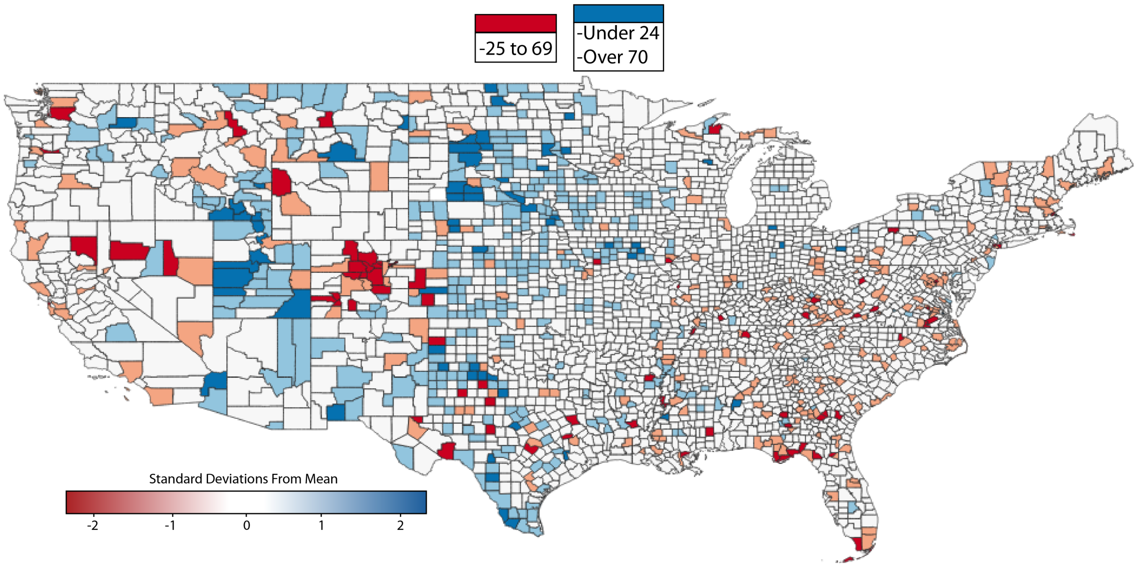
Figure 3.84: Age: Component 2 (2016)



The third component is less geographically clear cut than the first two, though concentrations of working age people can be seen spread about, particularly around

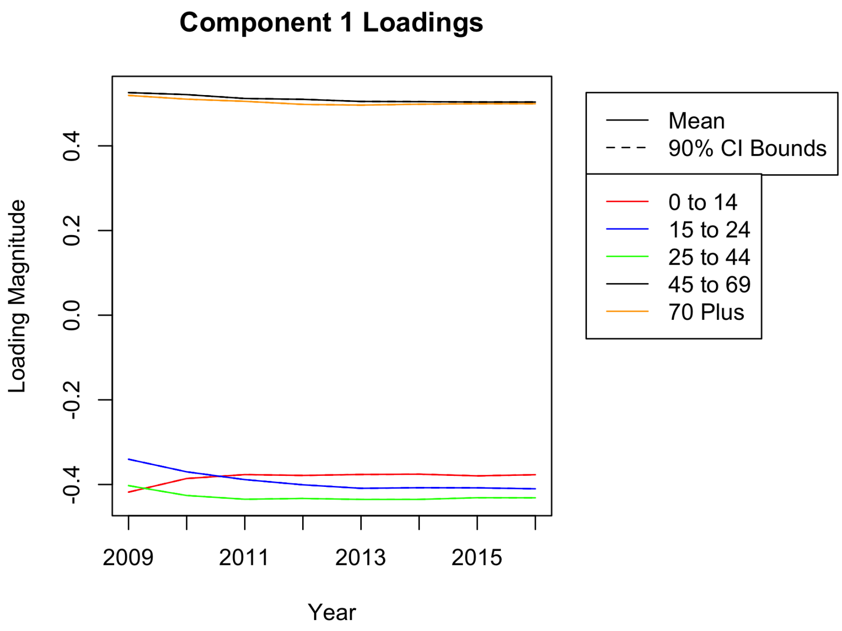
Denver, Colorado. Areas with more younger and more older people are concentrated around the central West and into parts of the Midwest, though Utah is stronger is this respect than most other areas.

Figure 3.85: Age: Component 3 (2016)



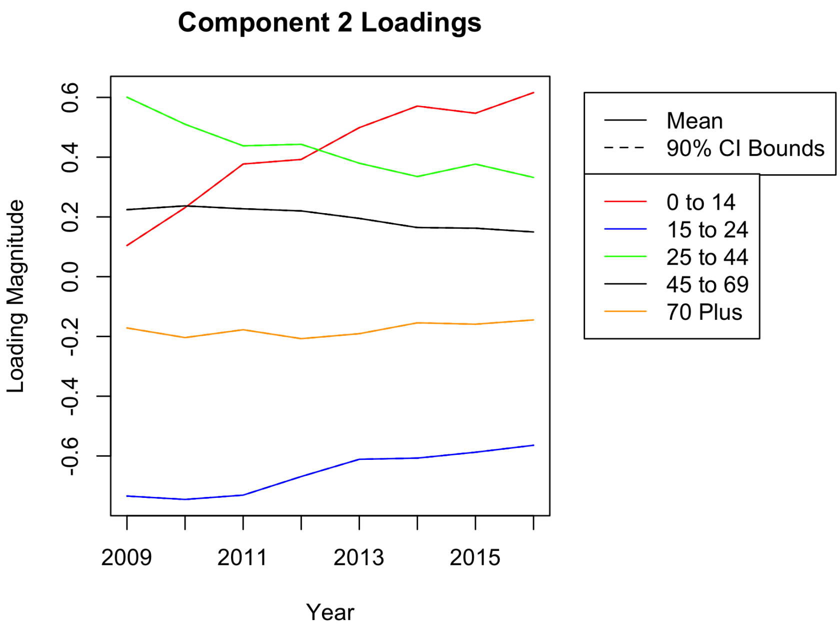
The plot below shows that the strongest age group component is extraordinarily robust with respect to time.

Figure 3.86: Age - Component 1: 2009 to 2016



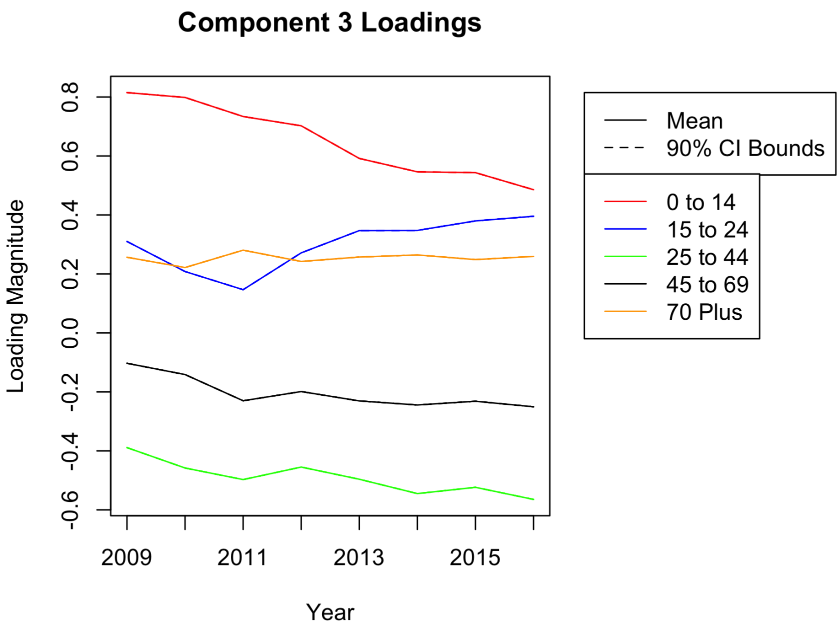
The second component compares less favorably to the first with respect to time however. The tendency of some counties towards more people between 15 and 24 is strong, but what this group is contrasted with changes over the study period. Earlier year show that the stronger trend is towards people in the next older age group in the positive direction, with this changing to a stronger emphasis on people aged under 14 as time progresses.

Figure 3.87: Age - Component 2: 2009 to 2016



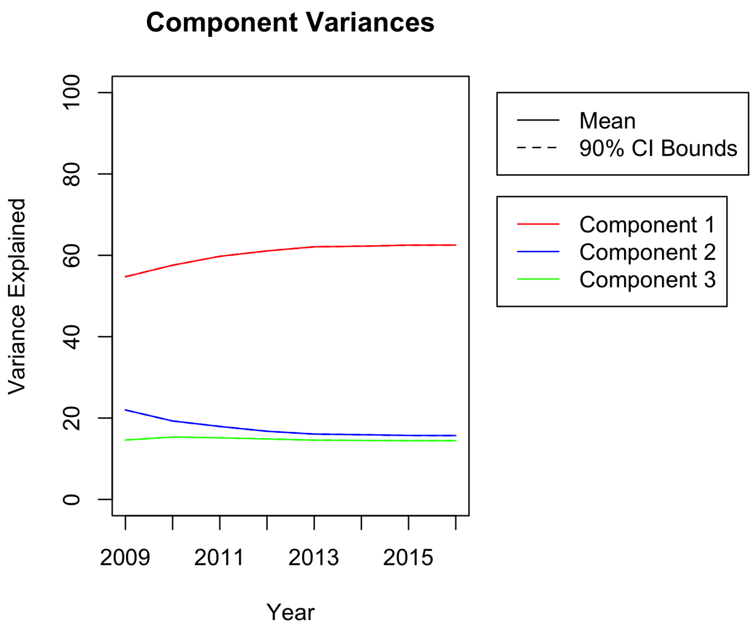
For component 3, the contributing strength of people under 14 is decreasing with time, though it remains important. On the other hand, age groups associated with the negative direction for this component are increasing in importance over the study period.

Figure 3.88: Age - Component 3: 2009 to 2016



The relative proportion of variances explained by each component are largely stable over time, though the distinction between counties the tend towards people over 45, versus under 45 is strengthening from 2009 to 2016.

Figure 3.89: Age - Proportion of Variance Explained: 2009 to 2016



Chapter 4

Regression Results

With the descriptive results in hand, focus can be turned to investigating labor force and employment growth. Labor force participation is the percentage of working-age people who are either working or actively looking for work. Employment figures are then calculated as a percent of people in the labor force, rather than as a percent of the entire working-age population.

Here we are estimating linear combinations (regression coefficients), of the linear combinations found by principal component analysis. This is not standard, but is justified through increased overall interpretability. Though, in theory, similar estimated coefficients could be found by examining the raw data, disentangling the multicollinearity effects present in the original Census variables is tremendously challenging. By using the constructed principal components instead of these variables, correlation-related issues are reduced, while enhancing clarity of understanding of the regression output. That is, a clearer picture of how different social effects are related to employment growth metrics can be seen when those social effects are well-specified principal components.

When estimating relationships between growth in these variables and the principal components above, it is important to have control variables for variation not otherwise accounted for. Each model estimated in this thesis has at least five such variables.

Population and population density differ greatly between counties, and may be an important driver of differing labor market trends. In addition, all information used in the principal component analyses are either percentages, or per capita measures, and so are stripped of their underlying population effects. Therefore, population and density are accounted for explicitly in each regression model.

A second possible major driver of job creation and labor force choices is the state of employment and residence. This could be caused by state-wide policies, or even just by general culture. To account for this possibility, state fixed effects are included in each regression model, and Alabama is selected as the default.

Finally, the absolute levels of labor force participation and unemployment are included because the absolute level could reasonably be expected to have a noteworthy effect on changes in that level. When estimating a model for sector-level employment changes, the level for that sector is included as well.

This leaves the basic model, shown in (4.1), where $\boldsymbol{Y}$ is one of the economic growth variables.

$$\begin{aligned}\boldsymbol{Y} &= \beta_0 + \beta_1\text{Population} + \beta_2\text{Population Density} + \beta_3\text{State} \\ &\quad +\beta_4\text{Unemployment} + \beta_5\text{Labor Force Participation} \\ &\quad (+\beta_6\text{Sector Employment})\end{aligned} \tag{4.1}$$

The basic model (4.1) accounts for the majority of variation explained in the dependent variables, except for in the secondary and quaternary sectors.

Table 4.1: Variation Explained by the Basic Model

	Basic Model - Adjusted-R^2	**Average Final Model Adjusted-R^2**
Employment	8.04%	9.29%
Labor Force Participation	3.89%	4.76%
Primary Sector	10.56%	11.54%
Secondary Sector	1.76%	3.19%
Tertiary Sector	4.45%	5.83%
Quaternary Sector	1.75%	4.00%

The smallest possible model used is the basic model (Equation (4.1)) with a single principal component from the previous chapter, see (4.2).

$$\begin{aligned} \boldsymbol{Y} &= \beta_0 + \beta_1 \text{Principal Component}_i + \beta_2 \text{Population} \\ &\quad + \beta_3 \text{Population Density} + \beta_4 \text{State} + \beta_5 \text{Unemployment} \\ &\quad + \beta_6 \text{Labor Force Participation } (+ \beta_7 \text{Sector Employment}) \end{aligned} \tag{4.2}$$

Additionally, other investigations into employment growth indicate that company size may be an important factor [16, 21]. These effects are clearly present in the following analyses, with only seven of the sixty regression models not benefiting from their presence. This benefit comes in terms of higher adjusted-R^2 values, more accurate principal component coefficients, and statistical significance of the variables in their own right. These coefficients are not evaluated unto themselves, instead serving as a tool to better estimate the relationship between previously identified principal components and employment growth metrics, however their importance cannot be denied.

Company size variables are created by the Census, and come in nine levels, shown below. This work incorporates nine for economy-wide firm sizes, as well as splitting the numbers out by sector. This gives a total of 54 company size variables.

Table 4.2: Ranges of Company Size Variables

Establishments with 1 to 4 employees	Establishments with 5 to 9 employees
Establishments with 10 to 19 employees	Establishments with 20 to 49 employees
Establishments with 50 to 99 employees	Establishments with 100 to 249 employees
Establishments with 250 to 499 employees	Establishments with 500 to 999 employees
Establishments with 1,000 employees or more	

To test these variables for importance, forward-stepwise model selection was employed. Each model is checked for prediction accuracy using cross-validation. Including more variables is not desirable in and of itself, so the smallest possible model is chosen once prediction error is accounted for.

Stepwise model selection does have shortcomings. Primarily, it may miss a collection of variables that is better in aggregate, even though the inclusion of the individual variables is not the most optimal at each step. A more robust approach is the 'best subset' selection approach, however, this technique is not used here due to the sheer complexity of running and checking the 2^{54} possible models for each of the examined principal components.

Finally, after each individual principal component model has been tested, a model containing all principal components is run. This model uses the same base model as above, and is referred to as the 'full model'. This model helps corroborate the significance of component-level effects found in individual models. These results are considered more robust, and are also discussed, though their confirmation does not preclude significance found in the single-component regression analyses.

4.0.1 Employment Growth

Table 4.3: Employment Growth: Significant Regression Coefficients

Component		**Coefficient**	**Adjusted-R^2**	**Business Effects**
Industry Sector	Comp.1	-0.067	9.34%	Y
	Comp.2	-0.183	10.49%	Y
	Comp.3	-0.049	8.99%	Y
Income	Comp.2	-0.091	9.41%	Y
Place of Employment	Comp.1	-0.083	9.27%	Y
	Comp.2	-0.051	8.94%	Y
	Comp.3	-0.054	9.07%	Y
Commute Time	Comp.1	0.088	9.35%	Y
Weeks Worked	Comp.1	0.163	9.22%	Y
	Comp.2	-0.048	8.81%	Y
Age	Comp.3	0.087	9.28%	Y

Table 4.4: Employment Growth: Principal Component Loadings for Significant Regression Components

Component	**Positive Loadings**	**Negative Loadings**
Industry Sector, Comp.1	Quaternary	Primary Secondary Tertiary
Industry Sector, Comp.2	Primary	Secondary Tertiary
Industry Sector, Comp.3	Tertiary	Secondary
Income, Comp.2	Less Than $25k More Than $100k	$25k to $50k $50k to $100k
Place of Employment, Comp.1	Worked in county of residence Worked in city of residence	Worked outside county of residence Not living in a city
Place of Employment, Comp.2	Worked in county of residence Not living in a city	Worked outside county of residence Worked outside city of residence
Place of Employment, Comp.3	Worked outside of state of residence	
Commute Time, Comp.1	30 to 44 minutes 45 or more minutes	Less than 10 minutes 10 to 14 minutes
Weeks Worked, Comp.1	50 to 52 40 to 49 1 to 39	Did not work
Weeks Worked, Comp.2	50 to 52	40 to 49 1 to 39 Did not work
Age, Comp.3	Under 14 15 to 24 Over 70	25 to 44 45 to 69

Each component tested is an axis with a negative direction and a positive direction. This means that all statistically significant results must be interpreted as having both a negative and a positive effect on the examined growth rate, depending on the sign of the coefficient and the portion of this axis being considered.

Table 4.3 for employment growth results shows a negative coefficient for each of the industry sectors. The principal component loadings shown in Table 4.4 show that there are positive loadings for all economic sectors but the secondary. The secondary sector is on the negative side for each component as well, so this is a sign that secondary industries, that is, construction and manufacturing, were growing in 2016-2017.

These analyses also find that middle income counties have higher job growth, as do counties where a disproportionate percentage of people must cross county-lines in their daily commutes. The latter may indicate growth in commuter suburbs, or so-called 'bedroom communities'. This is reinforced by the analysis of the first commuting times principal component. The positive coefficient here indicates that longer commute times are associated with overall increases in employment.

The third principal component for employment location data is strictly for people who cross state-lines to get to work, and the negative estimated coefficient indicates that this type of commute is not increasing in prevalence in 2016. It seems unlikely though that border communities are penalized for their proximity to other states. Instead, this may indicate people moving their places of residence across state-lines, thereby avoiding having to make the trip on a daily basis. Or perhaps businesses and employees do not need to look as far afield for opportunities as the general economy continues to grow.

Not all of the jobs created are full time, as indicated by the weeks worked principal component, and the age data indicate that many new employment positions are being taken up by younger people or possibly by former retirees (people over the age of 70). Together these may hint at a possible broader malaise in the economy, even as overall employment increases.

When these results are compared to a regression analysis containing all possible principal components, significance is preserved in the following cases:

Table 4.5: Employment Growth: Regression Coefficients Agreeing With Full Model

Components		Sign Change
Industry Sector	Comp.1	N
	Comp.2	N
	Comp.3	N
Commute Time	Comp.1	N
Weeks Worked	Comp.1	N
	Comp.2	Y
Age	Comp.3	N

This draws attention to employment growth in counties with more secondary sector industry and longer commutes, as well as to counties with more families on average (Age Component 3). The change in estimated regression coefficient sign for the second Weeks Worked component indicates that perhaps even less emphasis should be put on the belief that the jobs being created in 2016 are year-round positions (Weeks Worked Component 2), as this component might not retain significance in the face of deeper scrutiny.

4.0.2 Labor Force Participation Growth

Table 4.6: Labor Force Participation Growth: Significant Regression Coefficients

Component		Coefficient	Adjusted-R^2	Business Effects
Industry Sector	Comp.2	-0.063	4.02%	
Place of Employment	Comp.3	-0.048	4.07%	
Weeks Worked	Comp.2	-0.047	4.80%	Y
Households and Families	Comp.1	0.067	4.85%	Y
	Comp.2	0.050	4.84%	Y
Marital Status	Comp.2	0.070	4.15%	
Age	Comp.1	-0.100	5.34%	Y
	Comp.3	0.129	6.05%	Y

Table 4.7: Labor Force Participation Growth: Principal Component Loadings for Significant Regression Components

Component	**Positive Loadings**	**Negative Loadings**
Industry Sector, Comp.2	Primary	Secondary Tertiary
Place of Employment, Comp.3	Worked outside of state of residence	
Weeks Worked, Comp.2	50 to 52	40 to 49 1 to 39 Did not work
Households and Families, Comp.1	Average Household Size Households With Children Renter-Occupied Units	Owner-Occupied Units Single Unit Structures
Households and Families, Comp.2	Average Household Size Households With Children Single Unit Structures	Mobile Home Units
Marital Status, Comp.2	Never Married	Widowed Divorced Separated
Age, Comp.1	44 and Under	45 and Over
Age, Comp.3	14 and Under 15 to 24 70 and Over	25 to 44 45 to 69

Labor force participation dropped in communities with high proportions of people crossing state-lines. This mirrors the employment growth numbers seen previously, though again it is not clear if this is a drop in absolute magnitude, or a trend towards relocation. Also similar to the employment analysis, participation in the labor force increased for people aged between 15 and 24, or over 70. The first age component, on the other hand, finds an uptick in labor force participation amongst people over 45. The strength of this coefficient and model are less than for the third age analysis, but the component itself is stronger, explaining 62.5% of total age data variation, as opposed to the only 14.8% explained by component three. This makes the model for the first component the stronger net effect.

Though it is a smaller overall effect, the second age component covers people aged 14 and under, and this may by a sign that labor force participation is increasing in counties with more families. This is supported by the two households and families coefficients. Both of these have positive coefficients with respect to labor force participation growth, and both on average find larger households with children.

Finally, the negative component for weeks worked seems to indicate that people are who entering the labor force are not finding full-time work.

Table 4.8: Labor Force Participation Growth: Regression Coefficients Agreeing With Full Model

Components		**Sign Change**
Industry Sector	Comp.2	N
Weeks Worked	Comp.2	Y
Households and Families	Comp.1	N
Age	Comp.3	N

As with employment growth, the sign changed for the second component of Weeks Worked data. This means that there is still possibly some ambiguity as to if those entering the workforce are finding year-round work or not. Taken together, these two sign changes for this component indicate that further investigation is necessary to determine if job creation and labor force participation growth is skewed more towards year-round- or part-time work.

Otherwise, the components in Table 4.8 emphasize that labor force growth is concentrated in areas with more secondary and tertiary employment, as well as in areas averaging higher proportions of families.

4.0.3 Primary Sector Employment Growth

Table 4.9: Primary Sector Growth: Significant Regression Coefficients

Component		**Coefficient**	**Adjusted-R^2**	**Business Effects**
Industry Sector	Comp.2	-0.00056	11.81%	Y
Income	Comp.1	-0.00092	12.15%	Y
	Comp.2	-0.00045	11.82%	Y
Bachelor's Degree Field	Comp.2	-0.00034	11.82%	Y
Commute Time	Comp.2	-0.00048	11.02%	
Weeks Worked	Comp.1	0.00097	10.90%	
Marital Status	Comp.1	-0.00051	11.59%	Y
Age	Comp.1	0.00040	10.75%	
	Comp.3	0.00065	12.02%	Y

Table 4.10: Primary Sector Growth: Principal Component Loadings for Significant Regression Components

Component	**Positive Loadings**	**Negative Loadings**
Industry Sector, Comp.2	Primary	Secondary Tertiary
Income, Comp.1	More Than $50k	Less Than $50k
Income, Comp.2	Less Than $25k More Than $100k	$25k to $50k $50k to $100k
Field of Bachelor's Degree, Comp.2	Business	Science and Engineering Education
Commute Time, Comp.2	10 to 14 minutes 15 to 29 minutes	Less than 10 minutes 45 or more minutes
Weeks Worked, Comp.1	50 to 52 40 to 49 1 to 39	Did not work
Marital Status, Comp.1	Never Married Separated	Married
Age, Comp.1	44 and Under	45 and Over
Age, Comp.3	14 and Under 15 to 24 70 and Over	25 to 44 45 to 69

All coefficients estimated for primary sector job growth are quite small, though interestingly, the adjusted-R^2 values are the highest across any of the five metrics examined.

Two of the strongest effects are for the income components. The negative signs mean that primary sector growth is more associated with counties that on average tend more towards lower income households than higher income, and towards a more equal income distribution as opposed to a less equal distribution.

The Bachelor's degree component indicates that job growth in this sector of the economy is more associated with science, engineering, and education degrees. This may be because of the more technical nature of the oil and mineral extraction industries covered by the primary sector. It may also be that counties with more of these jobs being created are predominantly rural, and so the relatively few people who are educated to the Bachelor's level are either technical staff, or teachers, hence the education degrees.

Interestingly, counties associated with primary sector job growth also tend more towards quite short- or quite long commutes. As discussed in the commute times section, this appears to be an indicator of relatively rural communities, as well as jobs in primary sector industries. Finally, while people tend to be younger in these counties, they also tend more towards being married.

It would be of interest to distinguish between mineral extraction and agricultural pursuits in these components to see which trends were accentuated with respect to one portion of primary sector endeavors or another.

Checking the components in Table 4.9 against the model containing all principal components shows that the strongest effects found are for primary job creation in areas without high levels of primary industry, and with a trend towards incomes averaging below $50,000 per year.

Table 4.11: Primary Sector Growth: Regression Coefficients Agreeing With Full Model

Components		**Sign Change**
Industry Sector	Comp.2	N
Income	Comp.1	N

4.0.4 Secondary Sector Employment Growth

Table 4.12: Secondary Sector Growth: Significant Regression Coefficients

Component		Coefficient	Adjusted-R^2	Business Effects
Industry Sector	Comp.1	-0.0012	3.24%	Y
	Comp.3	0.0008	3.19%	Y
Place of Employment	Comp.2	0.0006	2.08%	
Households and Families	Comp.2	0.0006	3.75%	Y
	Comp.3	0.0008	3.44%	Y
Marital Status	Comp.1	-0.0010	2.99%	Y
Age	Comp.2	0.0007	3.64%	Y

Table 4.13: Secondary Sector Growth: Principal Component Loadings for Significant Regression Components

Component	Positive Loadings	Negative Loadings
Industry Sector, Comp.1	Quaternary	Primary Secondary Tertiary
Industry Sector, Comp.3	Tertiary	Secondary
Place of Employment, Comp.2	Worked in county of residence Not living in a city	Worked outside county of residence Worked outside city of residence
Households and Families, Comp.2	Average Household Size Households With Children Single Unit Structures	Mobile Home Units
Households and Families, Comp.3	Average Household Size Mobile Home Units Owner-Occupied Units	Single Unit Structures Renter-Occupied Units
Marital Status, Comp.1	Never Married Separated	Married
Age, Comp.2	14 and Under	15 to 24

Similar to the primary sector, this analysis shows that the secondary sector is not growing in counties that already have a high proportion of jobs in secondary industries. Rather, the primary and tertiary sectors seem to be expanding in these areas instead.

In addition, like growth in labor force participation, secondary sector job growth appears to have some relationship with the presence of families, as indicated by the last four components in Table 4.13.

Table 4.14: Secondary Sector Growth: Regression Coefficients Agreeing With Full Model

Componenst		**Sign Change**
Industry Sector	Comp.1	Y
	Comp.3	N
Marital Status	Comp.1	N

On closer inspection, the trend towards growth for this sector being in areas where it is not already prevalent is maintained. Whether secondary sector industries are growing in areas with higher levels of quaternary activity, however, is less clear. Interestingly, the only other effect found both in the individual and full models is a tendency towards higher levels of marriage in counties experiencing increases in secondary sector employment.

4.0.5 Tertiary Sector Employment Growth

Table 4.15: Tertiary Sector Growth: Significant Regression Coefficients

Component		**Coefficient**	**Adjusted-R^2**	**Business Effects**
Industry Sector	Comp.1	-0.0009	5.77%	Y
Income	Comp.2	-0.0006	5.76%	Y
Educational Attainment	Comp.1	-0.0017	7.06%	Y
	Comp.2	-0.0007	5.79%	Y
Bachelor's Degree Field	Comp.1	-0.0015	6.38%	Y
	Comp.2	0.0009	5.94%	Y
	Comp.3	0.0006	5.71%	Y
Place of Employment	Comp.1	-0.0008	5.94%	Y
Commute Time	Comp.1	0.0006	5.70%	Y
	Comp.2	-0.0010	6.19%	Y
Weeks Worked	Comp.1	-0.0021	6.16%	Y
	Comp.2	0.0006	5.71%	Y
Households and Families	Comp.2	0.0009	6.03%	Y
	Comp.3	0.0008	5.78%	Y
Age	Comp.1	-0.0007	4.69%	Y
	Comp.2	0.0005	4.66%	Y

Table 4.16: Tertiary Sector Growth: Principal Component Loadings for Significant Regression Components

Component	**Positive Loadings**	**Negative Loadings**
Industry Sector, Comp.1	Quaternary	Primary Secondary Tertiary
Income, Comp.2	Less Than $25k More Than $100k	$25k to $50k $50k to $100k
Educational Attainment, Comp.1	Some College Bachelor's Degree Graduate Degree	Less Than High School High School
Educational Attainment, Comp.2	Less Than High School Graduate Degree	Some College
Field of Bachelor's Degree, Comp.1	Science and Engineering Arts, Humanities and Others	Science and Engineering Rltd Fields Education
Field of Bachelor's Degree, Comp.2	Business	Science and Engineering Education
Field of Bachelor's Degree, Comp.3	Science and Engineering Rltd Fields Arts, Humanities and Others	Education
Place of Employment, Comp.1	Worked in county of residence Worked in city of residence	Worked outside county of residence Not living in a city
Commute Time, Comp.1	30 to 44 minutes 45 or more minutes	Less than 10 minutes 10 to 14 minutes
Commute Time, Comp.2	10 to 14 minutes 15 to 29 minutes	Less than 10 minutes 45 or more minutes
Weeks Worked, Comp.1	50 to 52 40 to 49 1 to 39	Did not work
Weeks Worked, Comp.2	50 to 52	40 to 49 1 to 39 Did not work

Table 4.17: Tertiary Sector Growth: Principal Component Loadings for Significant Regression Components (Continued)

Component	Positive Loadings	Negative Loadings
Households and Families, Comp.2	Average Household Size Households With Children Single Unit Structures	Mobile Home Units
Households and Families, Comp.3	Average Household Size Mobile Home Units Owner-Occupied Units	Single Unit Structures Renter-Occupied Units
Age, Comp.1	44 and Under	45 and Over
Age, Comp.2	14 and Under	15 to 24

Tertiary employment growth appears to be more oriented towards counties with middle incomes, and on average lower levels of education. The first educational component explains nearly 60% of all variation in the education data, and has the highest adjusted-R^2 across all tertiary sector regression models. This is a strong effect, and it is therefore striking that this analysis shows job growth for counties broadly defined by people with a high school education or less. Possibly this reflects lower education requirements in some tertiary industries. This is worthy of further investigation. The second education component calls attention to counties trending towards more people with some college, but in both cases Bachelor's degrees and higher are not defining features. These trends are sustained in regression modeling including all principal components simultaneously (Table 4.18).

For people that do have Bachelor's degrees in counties with increases in tertiary employment, the net result across the three components describing their fields is a tendency towards science and engineering related fields, as well as business degrees. These counties also appear to be characterized by moderate to long commutes across city- and county-lines.

For work status, these areas are notably defined by having more people who do not work at all, followed by a trend towards people working full time. This dichotomy seems to indicate an unsaturated employment market, perhaps one of the causes of the tertiary job growth being investigated. Finally, the counties found have on average larger households, sometimes with children, but otherwise generally tending towards older generations.

It is of interest to note that so many more principal components seem to be related to job growth in the tertiary industries than to growth in other areas. This sector covers basic services, as well as trade, so this may merely be because these industries are nearly ubiquitous, and so appear alongside all of the other variation present in American society. Even still, a deeper examination is called for.

Table 4.18: Tertiary Sector Growth: Regression Coefficients Agreeing With Full Model

Components		**Sign Change**
Educational Attainment	Comp.1	N
	Comp.2	N
Weeks Worked	Comp.1	Y
Households and Families	Comp.3	Y

4.0.6 Quaternary Sector Employment Growth

Table 4.19: Quaternary Sector Growth: Significant Regression Coefficients

Component		**Coefficient**	**Adjusted-R^2**	**Business Effects**
Industry Sector	Comp.3	0.0011	3.98%	Y
Educational Attainment	Comp.1	0.0027	4.55%	Y
Place of Employment	Comp.1	0.0009	3.90%	Y
	Comp.3	0.0007	3.87%	Y
Commute Time	Comp.2	0.0016	4.40%	Y
Households and Families	Comp.2	-0.0008	3.87%	Y
	Comp.3	-0.0008	3.80%	Y
Age	Comp.2	-0.0007	3.86%	Y
	Comp.3	-0.0008	3.80%	Y

Table 4.20: Quaternary Sector Growth: Principal Component Loadings for Significant Regression Components

Component	**Positive Loadings**	**Negative Loadings**
Industry Sector, Comp.3	Tertiary	Secondary
Educational Attainment, Comp.1	Some College Bachelor's Degree Graduate Degree	Less Than High School High School
Place of Employment, Comp.1	Worked in county of residence Worked in city of residence	Worked outside county of residence Not living in a city
Place of Employment, Comp.3	Worked outside of state of residence	
Commute Time, Comp.2	10 to 14 minutes 15 to 29 minutes	Less than 10 minutes 45 or more minutes
Households and Families, Comp.2	Average Household Size Households With Children Single Unit Structures	Mobile Home Units
Households and Families, Comp.3	Average Household Size Mobile Home Units Owner-Occupied Units	Single Unit Structures Renter-Occupied Units
Age, Comp.2	14 and Under	15 to 24
Age, Comp.3	14 and Under 15 to 24 70 and Over	25 to 44 45 to 69

The quaternary sector is represents the majority of employment in the United States. As such, continued job growth in this area is a significant driver of economic growth in general.

These new jobs are clearly tied to higher levels of education. Otherwise though, they seem to be in counties that are not particularly remarkable. The age components indicate that counties with new jobs in this sector tend to have more people aged 15 to 69, which is not valuable information when discussing the workforce. Similarly, the commute times indicated are in the middle of the possible range, and while working in the same city as ones residence is not particularly common, working in the same county certainly is (Figure 3.36).

Regressing the full set of principal components and other control variables on quaternary sector growth (Table 4.21) reinforces the relationship with higher levels of education. The other two effects found in both models are not particularly surprising, as recorded previously. This includes the clear association of tertiary employment, as most counties are expected to have around 20% of total employment in these areas, see Figure 3.1, and middling time commutes.

The components found to have a significant relationship mostly are also second or third tier components, meaning that they explain less variance than many of the previously discussed components. Likely both of these trends are due to the fact that most people in the United States are employed in quaternary industries. This means that all of the variety of the country is present when examining this industry, and so likely is present when examining its' growth as well.

As this sector of the economy is the most important to economic well-being in America, growth in this sector deserves an in-depth examination all its' own in future work.

Table 4.21: Quaternary Sector Growth: Regression Coefficients Agreeing With Full Model

Components		**Sign Change**
Industry Sector	Comp.3	N
Educational Attainment	Comp.1	N
Commute Time	Comp.2	N

Chapter 5

Discussion

The combination of principal component analysis (PCA) and linear regression is a powerful tool when exploring these data sets. Much structure was revealed in U.S. Census data that could not be seen without the multivariate perspective provided by PCA. These structures were also greatly simplified by way of their principal component deconstruction, increasing the confidence in, and value of, the estimated regression coefficients. This level of surety would have been challenging to achieve with the original, much more highly correlated, variables.

This analysis is strengthened by the greatly enhanced interpretability provided by PCA, as well as by the clear delineation of revealed groupings within the socio-economic and demographic variables. Together, these significantly clarify the results of linear regression analysis.

In addition, PCA reveals trends and patterns in a glance when plotted geographically. In some instances, regional distinctions are clearly evident. In others, those regions are still present, but sub-regional trends emerge that show connections that transcend physical geography.

Some areas of the country appear to be suffering in one respect, but are flourishing in another. Some counties find themselves consistently downtrodden, while others appear to be doing phenomenally well. Most counties must generally be average, by definition, but many then individually standout in unexpected ways. If no further lessons are taken from these analyses, the power of principal component analysis alone is evident.

But these components can do more than just describe patterns. Though no causal inference is assured, there are strong links between these social, demographic, and economic components and rates of job creation and labor force participation. These holy grails of economic and policy research are not associated solely with the firms and government policy, but exist in the complex morass of a shifting and evolving society. Establishing causality is, again, not trivial, but the results here show that this pursuit is not without hope.

Chapter 6

Appendix

6.1 Preliminary Principal Component Analysis

6.1.1 Age Preliminary PCA

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**
Standard deviation	2.7205	1.5065	1.2850	1.0869
Proportion of Variance	0.411	0.126	0.092	0.066
Cumulative Proportion	0.411	0.537	0.629	0.695

Table 6.1: Age - Preliminary Principal Component Analysis

	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**
Under 5 years	-0.247	-0.195	-0.337	-0.202
5 to 9 years	-0.207	-0.111	-0.475	-0.113
10 to 14 years	-0.158		-0.569	0.139
15 to 19 years	-0.185	-0.27		0.575
20 to 24 years	-0.22	-0.182	0.441	0.258
25 to 29 years	-0.267		0.312	-0.292
30 to 34 years	-0.255	0.139	0.152	-0.445
35 to 39 years	-0.2	0.241		-0.304
40 to 44 years	-0.11	0.428		0.148
45 to 49 years		0.514		0.192
50 to 54 years	0.167	0.39		
55 to 59 years	0.258	0.195		
60 to 64 years	0.298			
65 to 69 years	0.317			
70 to 74 years	0.312			
75 to 79 years	0.296	-0.153		-0.102
80 to 84 years	0.276	-0.203		-0.173
85 years and over	0.232	-0.225		-0.209

6.1.2 Weeks Worked Preliminary PCA

Table 6.2: Weeks Worked - Preliminary Principal Component Analysis

	Comp.1	**Comp.2**	**Comp.3**
Worked 50 to 52 weeks	-0.252	0.657	0.198
Worked 48 to 49 weeks	-0.321		-0.831
Worked 40 to 47 weeks	-0.46		-0.185
Worked 27 to 39 weeks	-0.42	-0.236	
Worked 14 to 26 weeks	-0.359	-0.4	0.185
Worked 1 to 13 weeks	-0.307	-0.39	0.399
Did not work	0.471	-0.448	-0.202

	Comp.1	**Comp.2**	**Comp.3**
Standard deviation	1.6264	1.3402	0.9262
Proportion of Variance	0.378	0.257	0.123
Cumulative Proportion	0.378	0.635	0.757

6.1.3 Income Preliminary PCA

Table 6.3: Income - Preliminary Principal Component Analysis

	Comp.1	**Comp.2**	**Comp.3**
Less than $10,000	-0.328	0.322	-0.101
$10,000 to $14,999	-0.353	0.165	-0.156
$15,000 to $24,999	-0.37		-0.106
$25,000 to $34,999	-0.306	-0.196	
$35,000 to $49,999	-0.176	-0.527	0.732
$50,000 to $74,999	0.144	-0.608	-0.516
$75,000 to $99,999	0.324	-0.223	-0.158
$100,000 to $149,999	0.384		
$150,000 to $199,999	0.359	0.223	0.16
$200,000 or more	0.323	0.274	0.306

	Comp.1	**Comp.2**	**Comp.3**
Standard deviation	2.3735	1.2598	0.7794
Proportion of Variance	0.564	0.159	0.061
Cumulative Proportion	0.564	0.722	0.783

6.1.4 Educational Attainment Preliminary PCA

Table 6.4: Educational Attainment - Preliminary Principal Component Analysis

	Comp.1	**Comp.2**	**Comp.3**
Less than 9th grade	-0.379	0.343	0.326
9th to 12th grade, no diploma	-0.461	0.129	0.233
High school graduate (includes equivalency)	-0.371	-0.339	-0.528
Some college, no degree	0.194	-0.454	0.72
Associate's degree	0.271	-0.526	-0.145
Bachelor's degree	0.487	0.226	
Graduate or professional degree	0.397	0.465	-0.141

	Comp.1	**Comp.2**	**Comp.3**
Standard deviation	1.8645	1.1761	0.9851
Proportion of Variance	0.497	0.198	0.139
Cumulative Proportion	0.497	0.694	0.833

6.1.5 Commute Times Preliminary PCA

Table 6.5: Commute Times - Preliminary Principal Component Analysis

	Comp.1	**Comp.2**	**Comp.3**
Less than 10 minutes	-0.417	0.262	-0.307
10 to 14 minutes	-0.305	-0.329	0.356
15 to 19 minutes		-0.55	0.358
20 to 24 minutes	0.214	-0.515	
25 to 29 minutes	0.329	-0.3	-0.377
30 to 34 minutes	0.411		-0.236
35 to 44 minutes	0.423	0.122	
45 to 59 minutes	0.399	0.237	0.28
60 or more minutes	0.264	0.296	0.602

	Comp.1	**Comp.2**	**Comp.3**
Standard deviation	1.9731	1.5243	0.9837
Proportion of Variance	0.433	0.258	0.108
Cumulative Proportion	0.433	0.691	0.799

6.2 Appendix - State Fixed Effects

State	Employment	Labor Force	Primary Sector
Arizona			
Arkansas	-0.331		
California	0.415		
Colorado			
Connecticut			
Delaware			
District of Columbia			
Florida	0.477		
Georgia	0.624	0.545	
Idaho			
Illinois			0.002
Indiana	0.31	0.348	
Iowa		0.407	
Kansas		0.555	
Kentucky			-0.003
Louisiana	-0.375		-0.002
Maine			
Maryland			
Massachusetts			
Michigan	0.465	0.375	
Minnesota			
Mississippi			
Missouri		0.29	
Montana		0.621	
Nebraska		0.497	
Nevada			
New Hampshire			
New Jersey			
New Mexico	-0.525	-0.559	
New York			
North Carolina	0.494	0.407	
North Dakota		0.727	

State	Secondary Sector	Tertiary Sector	Quaternary Sector
Arizona			
Arkansas	-0.005		
California		-0.003	
Colorado	0.004		
Connecticut			
Delaware			
District of Columbia			
Florida			
Georgia			
Idaho			
Illinois			
Indiana			
Iowa			
Kansas			
Kentucky		-0.004	0.007
Louisiana			
Maine			
Maryland			
Massachusetts			
Michigan		-0.004	
Minnesota			
Mississippi			
Missouri			
Montana			
Nebraska			
Nevada			
New Hampshire		-0.006	
New Jersey			
New Mexico			
New York			
North Carolina			
North Dakota	0.006	-0.008	

State	Employment	Labor Force	Primary Sector
Ohio		0.31	
Oklahoma			
Oregon	0.6	0.388	
Pennsylvania		0.31	
Rhode Island			
South Carolina			
South Dakota		0.627	-0.005
Tennessee	0.389	0.364	
Texas		0.396	
Utah		0.45	-0.004
Vermont			
Virginia			
Washington			
West Virginia			-0.004
Wisconsin	0.415	0.431	
Wyoming			-0.005

State	Secondary Sector	Tertiary Sector	Quaternary Sector
Ohio			
Oklahoma			
Oregon			
Pennsylvania			
Rhode Island			
South Carolina			
South Dakota			
Tennessee			
Texas			
Utah			
Vermont			
Virginia			
Washington			
West Virginia			0.006
Wisconsin			
Wyoming		-0.004	0.008

6.3 Appendix - Component Correlation With Total Population

Table 6.6: Correlation Between Population and Principal Components

Data Set	**Component (2016)**	**Correlation With 2016 Population (%)**
Age	Comp. 1	-21.7%
	Comp. 2	-1.2%
	Comp. 3	-18.4%
Bachelor's Degree	Comp. 1	29.8%
	Comp. 2	19.0%
	Comp. 3	5.9%
	Comp. 4	6.4%
Commute Time	Comp. 1	23.7%
	Comp. 2	20.8%
Education Attainment	Comp. 1	25.5%
	Comp. 2	25.5%
Households and Families	Comp. 1	30.1%
	Comp. 2	12.2%
	Comp. 3	-14.9%
Income	Comp. 1	20.8%
	Comp. 2	21.8%
Industry Sector	Comp. 1	31.8%
	Comp. 2	-15.0%
	Comp. 3	3.7%
Marital Status	Comp. 1	20.0%
	Comp. 2	21.6%
Place Of Employment	Comp. 1	22.0%
	Comp. 2	-18.0%
	Comp. 3	4.8%
Weeks Worked	Comp. 1	6.8%
	Comp. 2	0.5%

6.4 Appendix - Regression Results Using All Principal Components

Table 6.7: Employment Growth - Regression With All Principal Components (Only Significant Coefficients Shown)

Data Set	Component	Coefficient
Age	Comp.3	0.089
CommuteTime	Comp.1	0.179
	Comp.2	0.067
Households	Comp.1	0.208
	Comp.3	-0.206
Income	Comp.1	0.114
Sector	Comp.1	-0.122
	Comp.2	-0.154
	Comp.3	-0.052
WeeksWorked	Comp.1	0.354
	Comp.2	0.107

Table 6.8: Labor Force Participation Growth - Regression With All Principal Components (Only Significant Coefficients Shown)

Data Set	Component	Coefficient
Age	Comp.3	0.091
CommuteTime	Comp.1	0.102
	Comp.2	0.062
Households	Comp.1	0.193
	Comp.3	-0.161
Income	Comp.1	0.117
Sector	Comp.2	-0.098
	Comp.3	-0.050
WeeksWorked	Comp.1	0.360
	Comp.2	0.100

Table 6.9: Primary Sector Growth - Regression With All Principal Components (Only Significant Coefficients Shown)

Data Set	**Component**	**Coefficient**
CommuteTime	Comp.1	-0.001
Households	Comp.1	-0.001
Income	Comp.1	-0.001
Sector	Comp.2	-0.001

Table 6.10: Secondary Sector Growth - Regression With All Principal Components (Only Significant Coefficients Shown)

Data Set	**Component**	**Coefficient**
Bachelors	Comp.3	0.000
CommuteTime	Comp.2	-0.001
Marital	Comp.1	-0.001
Sector	Comp.1	0.004
	Comp.2	0.003
	Comp.3	0.004

Table 6.11: Tertiary Sector Growth - Regression With All Principal Components (Only Significant Coefficients Shown)

Data Set	**Component**	**Coefficient**
Education	Comp.1	-0.004
	Comp.2	-0.001
Households	Comp.3	-0.001
Income	Comp.1	0.002
WeeksWorked	Comp.1	0.003

Table 6.12: Quaternary Sector Growth - Regression With All Principal Components (Only Significant Coefficients Shown)

Data Set	**Component**	**Coefficient**
Age	Comp.1	0.002
Bachelors	Comp.3	0.001
CommuteTime	Comp.2	0.001
Education	Comp.1	0.004
Sector	Comp.1	-0.005
	Comp.2	0.001
	Comp.3	0.002
WeeksWorked	Comp.1	-0.003

References

[1] (2009). Calculating American Community Survey (ACS) Reliability. http://help.arcgis.com/en/communityanalyst/apis/rest/reference/ACSVariables.html. Accessed: 2019-04-28.

[2] (2009). US Census: Field of Bachelor's Degree in the United States: 2009. https://www2.census.gov/library/publications/2012/acs/acs-18.pdf. Accessed: 2019-04-28.

[3] (2010). US Census: Understanding "Place" in Census Bureau Data Products. https://www.census.gov/content/dam/Census/data/developers/understandingplace.pdf. Accessed: 2019-04-28.

[4] (2011). United Kingdom Office for National Statistics: 2011 residential-based area classifications. https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications. Accessed: 2019-04-28.

[5] (2012). US Census - American FactFinder: Tables and Products. https://factfinder.census.gov/help/en/tables_and_products.htm. Accessed: 2019-04-28.

[6] (2017). US Census: American Community Survey Information Guide. https:

//www.census.gov/programs-surveys/acs/about/information-guide.html. Accessed: 2019-04-28.

[7] (2018). US Census: Distinguishing features of acs 1-year, 1-year supplemental, 3-year, and 5-year estimates. https://www.census.gov/programs-surveys/acs/guidance/estimates.html. Accessed: 2019-04-28.

[8] (2019). Environmental Systems Research Institute (ESRI): Tapestry Segmentation - Insights into America's Changing Population. https://www.esri.com/tapestry. Accessed: 2019-04-28.

[9] (2019). Parish Government Information. https://www.lpgov.org/page/ParishInfo. Accessed: 2019-04-28.

[10] (2019). US Census: American FactFinder. https://factfinder.census.gov. Accessed: 2019-04-28.

[11] (2019). US Census Geography Program - Glossary. https://www.census.gov/programs-surveys/geography/about/glossary.html. Accessed: 2019-04-28.

[12] Ball, L. M. (2014). Long-term damage from the great recession in oecd countries. Technical report, National Bureau of Economic Research.

[13] Borbely, J. M. (2011). Characteristics of displaced workers 2007–2009: a visual essay. *Monthly Labor Review*, 134(9):3–15.

[14] Carlino, G. A. and Mills, E. S. (1987). The determinants of county growth. *Journal of Regional Science*, 27(1):39–54.

[15] Carnevale, A. P., Smith, N., and Strohl, J. (2013). Recovery: Job growth and education requirements through 2020. *Georgetown University Center on Education and the Workforce*.

[16] Criscuolo, C., Gal, P. N., and Menon, C. (2014). The dynamics of employment growth: New evidence from 18 countries.

[17] Davis, S. J., Haltiwanger, J. C., Schuh, S., et al. (1998). Job creation and destruction. *MIT Press Books*, 1.

[18] Duca, J. V. (2013). Subprime mortgage crisis. *Federal Reserve History*, 22:22.

[19] Fuller, S. (2018). Using american community surveyestimates and margins of error. https://www.census.gov/content/dam/Census/programs-surveys/acs/guidance/training-presentations/20180418_MOE.pdf. Accessed: 2019-04-28.

[20] Grove, D. M. and Roberts, C. A. (1980). Principal component and cluster analysis of 185 large towns in england and wales. *Urban Studies*, 17(1):77–82.

[21] Henderson, J. and Weiler, S. (2010). Entrepreneurs and job growth: probing the boundaries of time and space. *Economic Development Quarterly*, 24(1):23–32.

[22] Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

[23] Kenessey, Z. (1987). The primary, secondary, tertiary and quaternary sectors of the economy. *Review of Income and Wealth*, 33(4):359–385.

[24] Mills, E. S. (1986). Metropolitan central city population and employment growth during the 1970s. *Prices, competition and equilibrium. Philip Allan, London*, pages 268–284.

[25] Partridge, M. D. and Rickman, D. S. (2007). Persistent pockets of extreme american poverty and job growth: Is there a place-based policy role? *Journal of Agricultural and Resource Economics*, pages 201–224.

[26] R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.