Examining County-Level Economic Performance Since the Great Recession

From: Eric LaRose
To: Professor Brodnax
Date: May 4, 2020

Executive Summary

This project considers factors that explain county-level economic performance since the Great Recession (2009-2016), as measured by percentage changes in median household income, total employment, and the unemployment rate. These factors are primarily county-level demographic and economic characteristics as measured at the height of the Great Recession, in 2009. Decision tree and linear regression models are applied to each dependent variable, separately, to find the factors that best explain each variable. Overall, both models appear to have limited ability for accurate evaluation and prediction, and are able to explain only a minor share of variation in each dependent variable. Furthermore, for a given dependent variable the two models generally disagree on which variables are most important. A major exception is that both models agree that unemployment rate in 2009 is by far the most important factor in predicting subsequent changes in the unemployment rate. From these results, we cannot conclude that there is a single factor, or set of factors, that is extremely important in explaining all three of these measures of economic performance.

Background

This project examines potential factors that explain county-level economic performance since the Great Recession, with a particular focus on which factors might be most important in explaining this economic performance. Overall, since the Great Recession the U.S. economy has grown steadily, adding approximately twenty million jobs between 2009 and 2020 ("All Employees, Total Nonfarm," 2020). Likewise, inflation-adjusted median household income grew by over ten percent between 2009 and 2018 ("Real Median Household Income in the United States," 2019).

While these national statistics paint a rosy picture of America's economic recovery, they obscure considerable regional variations. As Schneider (2019) documents, forty percent of job and wage growth since the Great Recession has occurred in just twenty metropolitan areas that make up only one-quarter of the nation's population, and only about nine percent of job growth has occurred in rural counties. Dozens of metropolitan areas still have fewer jobs than they did before the Great Recession (Sauter, 2018). These facts, and many others, paint a more nuanced picture of the country's economic recovery, in which so-called "superstar" cities primarily located on the coasts or in the South enjoy substantial economic gains (Schneider, 2019) while much of the rest country experiences economic stagnation.

Economic growth that is very unevenly distributed across space can have many wide-ranging consequences. For instance, previous research has shown that long unemployment spells increase the risk of suicide (Classen and Dunn, 2012) and may have helped to fuel the opioid crisis ravaging economically disaffected areas in Appalachia and the Rust Belt (Hollingsworth, Ruhm, and Simon, 2017). Additionally, manufacturing job losses in the Midwest may have helped elect Trump; residents of counties most exposed to import competition from China became more polarized in their political views (Autor, Dorn, Hanson, and Majlesi, 2017).

Given the evidence that America's economic gains since the Great Recession are unevenly distributed, and that this trend has far-reaching consequences, it is extremely important for policymakers to better understand the factors that explain differences in regional economic growth. A better understanding of these factors may help design policies that promote better economic outcomes in areas that have largely been left out of the past decade's economic expansion.

This, this project is primarily focused on inference – understanding how economic growth is affected by various factors, and which factors are potentially most important in explaining this growth. However, this project also considers prediction, in terms of how well these factors can, together, predict economic growth. In this project, a region is defined as a county, since counties are the smallest geographic units for which government agencies annually release precise data on various economic and demographic characteristics. There are currently 3,141 county-equivalent units in the United States, with an average population of slightly over 100,000 ("How Many Counties are in the United States?", n.d.).

While relevant county-level data is relatively easy to obtain, another challenge consists of determining economic performance. There is no commonly agreed-upon measure of economic performance, and even if there were, such a variable might not be directly measurable. Two common proxy variables for economic performance are income growth, as measured by either median household or per capita income, and employment growth (see, for instance, Cassidy (2018) and "Chart Book" (2020)). Another common measure is to look at changes in the unemployment rate (Cunningham, 2018).

If these three measures are very highly correlated, then the choice of which measure to use arguably does not matter much, as counties doing well by one of these three metrics will generally be doing well by the other two. However, if these measures are not highly correlated, then we should conduct our analyses with all three measures as possible dependent variables. Indeed, there is some justification to expect that employment growth, or changes in the unemployment growth, might be weakly correlated with income growth. Particularly, cities that have attracted an extremely disproportionate share of new high-paying jobs, such as San Francisco, New York, and San Jose, tend to have very strict zoning laws that greatly restrict population and employment growth (Hsieh and Moretti, 2019). In the following section, we show that these measures are in fact weakly correlated with each other.

Data

Description of Dependent Variables

As mentioned above, the three main dependent variables of interest are county-level income growth, employment growth, and changes in the unemployment rate. The reason we use changes in the unemployment rate, in addition to employment growth, is that employment growth is likely to be highly correlated with population growth, which is at least to some extent probably exogenous to local economic conditions. The unemployment rate reflects the percentage of people in the workforce who are able to find a job, and thus should be less susceptible to exogenous changes in population.

Because we look at county-level recovery since the Great Recession, we look at percentage changes in all of these variables by county from 2009, when unemployment peaked and the recession officially ended ("The Recession of 2007-2009," 2012), through 2016, the latest year on which data for all three of these variables is available. Of course, one major limitation of this data is that it does not include data on 2017, 2018, and 2019, effectively missing over one quarter of the economic recovery. It would be interesting to redo the analysis once all of this data is available several years in the future, to see if findings are robust to looking at the period from 2009 through 2019, rather than 2009 through 2016. All three of these variables are calculated as percentage changes (rather than absolute changes) from 2009 through 2016.

Data on employment growth can be calculated from the County Business Patterns (CBP) dataset provided by the Census Bureau ("CBP Datasets," n.d.). CBP data is available annually from 1964 through 2017, and in each year contains data on aggregate employment in each county, and county-level employment broken down by industry. One potential limitation of this data is that, for county-industry pairs with small employment values, some values are censored or infused with noise to protect the confidentiality of data ("Methodology," 2019). When values are censored, codes are provided that indicate the bins containing the true employment value (0-19, 20-49, and so on). When this is the case, we assign employment values as the midpoint of these bins, rounded to the nearest whole number (for instance, 10 is the rounded midpoint of the 0-19 bin).

Data on income growth, measured as either median household or per capita income, is available through the Census Bureau's American Community Survey, or ACS ("American Community Survey Data," 2019). The ACS randomly surveys a sample of households in each year and produces estimates of various demographic and economic characteristics at the county level. The ACS currently produces one-year and five-year estimates, where five-year estimates are averages of estimates over a five-year period. Because one-year estimates are only published for counties with at least 65,000 people, we must use the five-year estimates ("When to Use 1-year, 3-year, or 5-year Estimates," 2019). The most recent available is from the 2014-2018 5-year ACS, centered on 2016. We use the 2014-2018 5-year averages as 2016 data, but an obvious limitation of this data is that of course these are not estimates from a single year. For variables that have a linear trend over this five-year period, however, an average should be close to the "true" 2016 value. For this project, we measure income growth in terms of median household income, rather than income per capita, because income per capita can be highly right-tail skewed by high earners. Median household income gives a better measure of how the median or "typical" resident of a county is faring (Amadeo, 2019).

Data on changes in the unemployment rate comes from the Local Area Unemployment Statistics (LAUS) dataset provided by the Bureau of Labor Statistics, using county-level annual averages across months in a year from 2009 and 2016 ("Local Area Unemployment Statistics," n.d.). As with ACS data, data on the unemployment rate ("How the Government Measures Unemployment," 2015) comes from a random survey of households and is subject to both sampling error and measurement error ("Error Measurement," n.d.).

Description of Independent Variables

We include in our analysis a wide variety of county-level characteristics which could potentially explain local economic growth.

The first set of variables includes demographic and socioeconomic county-level characteristics from 2009, in levels. It is possible, for instance, that counties that were wealthier or more educated to begin with have had a stronger recovery from the Great Recession. From the ACS 2007-2011 5-year ACS, centered on 2009, we gather the following variables on each county: total population, the percent of the population aged 25 or over with at least a bachelor's degree, median household income, and the percent of the population in poverty. This data is collected using the National Historical Geographic Information System ("IPUMS NHGIS," n.d.)

provided by the University of Minnesota. This ACS data is subject to the same limitations described for growth in median household income; namely, these values are five-year averages rather than point estimates, and for counties with small populations, the margins of error can be quite large. Supplementing the data on population with data on counties' land area from the Census Bureau's Gazetteer Files ("Gazetteer Files," n.d.), we also calculate each county's population density in 2009. Additionally, from the BLS's LAUS dataset we include each county's average unemployment rate throughout 2009 ("Local Area Unemployment Statistics," n.d.). From the County Business Patterns dataset we gather the county's total employment in 2009, and additionally calculate the percentage of each county's employment in 2009 that is in the manufacturing sector, defined as NAICS code 31 ("CBP Datasets," n.d.). These data are subject to the same limitations described above in the description of dependent variables.

Additionally, we are interested in variables for several county-level policy characteristics. Potential policy variables of interest include local tax burdens, the quality of local education, and the stringency of local zoning regulations. However, reliable data on county-level policy characteristics is hard to find. For instance, data on county-level tax burdens is not readily available in part because most taxes are either set at the state or local level, such that tax burdens can vary across cities within a county. Data on quality of public schools at the school-level is available from GreatSchools, a nonprofit organization focused on providing education information ("School Census Data," n.d.), but counties can contain hundreds of individual schools and dozens of school districts, with substantial heterogeneity in school quality. Additionally, data on local zoning regulations is available from the Wharton Residential Land Use Regulation Index (WRLURI), provided by the University of Pennsylvania ("Wharton Residential Land Use Regulation Index (WRLURI)," n.d.). However, the WRLURI provides information on cities, rather than counties, and only covers about 2600 municipalities. That is, many counties have no cities featured in the WRLURI, and very few counties have all of their municipalities contained in the WRLURI.

Because we are not able to find quality data on county-level policy characteristics, we instead measure the percentage of votes that Barack Obama received in the 2008 presidential election in each county, which occurred during the Great Recession. This variable is meant as a rough proxy for each county's local policy environment, in that it is likely correlated with several local policy characteristics. For instance, more liberal counties tend to have stricter zoning laws (Florida, 2018), and more liberal counties and states tend to have higher property tax rates (Blankley, 2019).

Likewise, we also measure for each county the percent of votes received by Obama in that county's corresponding state, which is meant as a proxy for state-level policy environments. One obvious limitation of these proxy variables is that it is impossible to disentangle the effects of specific policies. Data on county-level and state-level vote shares comes from the MIT Election Lab ("MIT Election Lab," n.d.). In addition to this proxy, we are also able to obtain information on state-level tax burdens in 2009 from the Tax Foundation ("State-Local Tax Burden Rankings," 2016). This data measures the percentage of the average resident's income in each state that goes toward paying state-level taxes, including income and sales taxes.

Cleaning Steps

After an initial review of the data, we decided to exclude from the analysis any county with a population of under 25,000 in 2009. There are several reasons for this. First, in small counties, relatively low changes in absolute numbers in employment (or other variables) can correspond to extremely high percentage changes. For instance, many very rural counties in states such as North Dakota, South Dakota, and Texas have gained a few hundred or thousand jobs during this decade's oil boom, which corresponds to percentage growth of over 150%. Second, small counties tend to have substantially higher margins of error in ACS data, and are also much more likely to have censored or imputed employment values in CBP data. While the 25,000 threshold is somewhat arbitrary, it should exclude almost all counties suffering from these issues, and the resulting dataset still contains slightly over half of all counties, at 1,603.

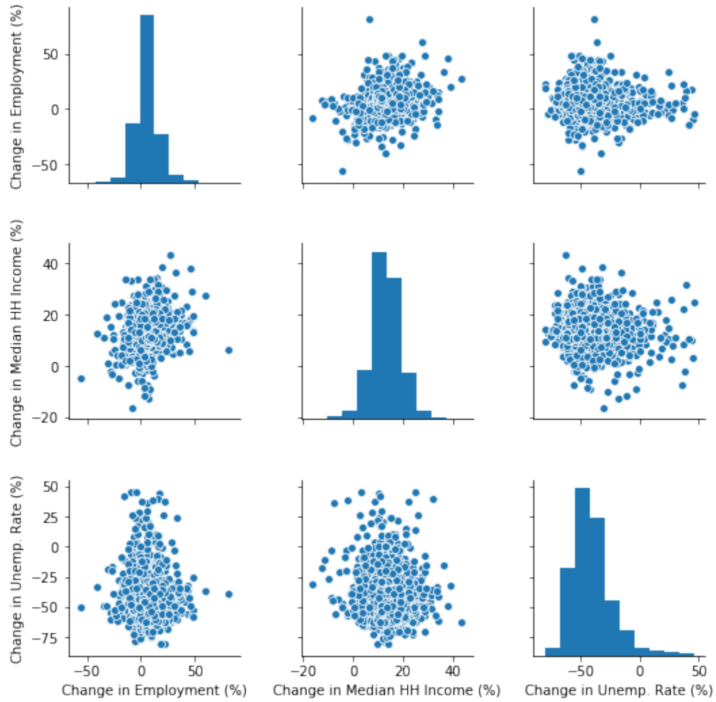
Additionally, numerical summaries of the data reveal that total population, total employment, and population density in 2009 are all highly right-tail skewed, but logged versions of these variables appear roughly normally distributed. Thus, we use the logged variables, rather than these variables in levels, in this report.

Summary of Dependent Variables

First, we investigate the correlations between our three potential dependent variables. As mentioned above, if these variables are all very highly correlated with each other, then it may be sufficient to just use one variable to measure county-level economic performance. However, this is not the case. The correlation of employment growth and median household income growth by county is just 0.23, the correlation of income growth and the percentage change in the unemployment rate is just -0.02, and the correlation of employment growth and the change in the unemployment rate is -0.07. The plot below shows scatterplot for each of these possible combinations of variables below and confirms that the relationships between each of our three variables are surprisingly weak.

Changes in Income, Employment, and Unemployment are Weakly Correlated

Across Counties with at Least 25,000 People, between 2009 and 2016



Additionally, the regression table below shows the results of regression in which we regress each of these three variables, one at a time, on the other two variables. Each dependent variable appears in the columns, while the rows indicate coefficient estimates and stars denote statistical significance. The R-squared values at the bottom indicate the percentage of variation in the dependent variable that can be explained by variation in the independent variable. As we can see, a given pair of these three variables can explain, at most, six percent of the variation in the other variable. This provides further evidence that these three variables are quite weakly related. Thus, we conduct each of the analyses in this paper three separate times, once each using each of these three possible dependent variables.

	Change in Employment (%)	**Change in Median HH Income (%)**	**Change in Unemp. Rate (%)**
Change in Employment (%)		0.13***	-0.11***
Change in Median HH Income (%)	0.39***		-0.02
Change in Unemp. Rate (%)	-0.04***	0	
Constant	-1.14	12.46***	-40.23***
N	1603	1603	1603
R^2	0.06	0.05	0.01

* p<.1, ** p<.05, ***p<.01

Summary of Independent Variables

The table below provides summary statistics of all the relevant independent variables that are used in the analysis. All variables are from 2009, except for the variables containing vote results from the 2008 presidential election. All of these variables appear to be roughly normally distributed, in that their means lie fairly close to their medians.

	Median Household Income ($)	**Poverty Rate (%)**	**Pop. w/ College Degree (%)**	**Share of Employment in Manufacturing (%)**	**Unemployment Rate (%)**	**Percent of Votes for Obama in 2008**	**Percent of State's Votes for Obama in 2008**	**State's Tax Burden Rate**	**Log(County Population)**	**Log(County Employment)**	**Log(County Population Density)**
County	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0
Mean	48830.23	15.17	22.27	15.7	9.56	44.7	48.03	10.04	11.38	10.14	4.92
Standard Deviation	12802.09	5.57	9.65	10.47	3.04	12.98	11.75	1.22	1.04	1.24	1.33
Minimum	22353.0	3.45	7.19	0.27	2.55	12.46	2.07	7.0	10.13	6.95	0.88
25th Percentile	40466.0	11.37	15.03	7.68	7.47	35.4	43.0	9.3	10.57	9.24	4.08

Median	45992.0	14.85	19.67	13.4	9.14	44.09	49.95	10.0	11.08	9.83	4.68
75th Percentile	53894.0	18.27	27.35	21.55	11.32	53.18	55.15	10.5	11.96	10.8	5.62
Maximum	120096.0	40.38	70.66	68.34	27.37	92.46	92.46	12.8	16.1	15.12	11.15

Methodology

Overall, in this project we are primarily interested in inference, and more specifically in finding which of the potential explanatory variables above are most important in predicting how counties have fared economically since the Great Recession. Therefore, we need to use analytical techniques that allow us to assess some measure of importance for each variable.

Two techniques that allow us to do this are linear regression and decision trees. Linear regression is a parametric approach, meaning that it requires us to assume a specific functional form. In contrast, decision trees are a non-parametric approach, in that they do not require any assumptions of functional form. These techniques are described in more detail below.

It is important to note that, when fitting a linear regression model and a decision tree to our data, we must first split our data up, randomly, into two separate sets. The first set is called the training set and contains 80% of all observations; we use the training set to fit each model. We then use the remaining 20% of all observations as the test data; once we have fit each model to our training data, we assess model accuracy using our test data. It is important to assess model fit on the "unseen" test data, rather than the training data, because there is the possibility of overfitting in the training data. Overfitting means that our model is fitting the noise in the data, and thus will have high accuracy on the training data but low training when applied to new data, where such noise will not exist.

We consider six total models in this paper, as we apply linear regression and a decision tree to each of our three dependent variables. We evaluate each model on two metrics: the R-squared value and the root mean squared error. The R-squared value indicates the fraction of variance in the dependent variable of interest that is explained by the model (James et al., 2017). The root mean squared error (RMSE) is, as the name implies, the square root of the mean squared error. The RMSE can be roughly interpreted as an approximation of the average distance of estimated values from observed values, in units of the dependent variable (Bailey, 2016).

Description of Linear Regression

The general equation for a multivariate linear regression is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \epsilon_i.$$

In this equation, i indexes observations, Y refers to the dependent variable of interest, β_0 refers to an intercept, and ϵ_i refers to an error term. Essentially, linear regression assumes that some dependent variable of interest can be approximated as a linear combination of a prespecified set of n independent variables, plus some error. Here, X_{1i} denotes the value of variable 1 for observation i, and so on. Linear regression chooses the intercept and set of coefficient estimates, $\beta_1, \ldots, \beta_n$ to minimize the sum of squared errors, ϵ_i^2, from the regression. Predicted values of the dependent or response variable can then be calculated by multiplying the estimated coefficients by the observed values for each variable (plus the constant) and summing.

Each coefficient estimate comes with a range of uncertainty, as these are merely estimates of the "true" value of the parameter. Hypothesis testing can be used to calculate the probabilities, or p-values, of observing coefficients as large or larger in magnitude given that the "true" value of a given β is zero. That is, if a coefficient estimate equals 1, with a p-value of

0.05, then that means that if the true value of the coefficient were zero, we would have only a 5% chance of observing a coefficient with an absolute value of at least one. Throughout this paper, we conclude that a p-value below 0.05 is statistically significant, meaning that we conclude the estimated coefficient is nonzero. For coefficient estimates with p-values above 0.05, we say that we do not have sufficient evidence to conclude the true coefficient is nonzero.

A key strength of linear regression is interpretability. The expected effect of a one-unit increase in a variable X_n on the value of the outcome variable is simply β_n. For variables measured in the same units, it is thus very easy to compare the magnitudes of the estimated effects; along with statistical significance, this provides a measure of variable importance. In this paper, we standardize all independent variables and the response variable when conducting a linear regression, meaning that we subtract the variable's mean from each value and divide by its standard deviation. Thus, all variables in our regression have a mean of zero and standard deviation of one. This means that all coefficient estimates are comparable, even for variables ordinarily measured in very different units. Particularly, each coefficient denotes the standard-deviation increase in the response variable associated with a one standard-deviation increase in the coefficient's corresponding variable.

A drawback of linear regression is flexibility. As previously mentioned, linear regression requires us to assume a specific functional form of the data, namely that the response variable is linear in the predictors. If this assumption is far from reality, then linear regression will likely have a relatively poor fit on our test data (James et al., 2017). Because of linear regression's low flexibility, the method also suffers from relatively high bias. This means that linear regression suffers from relatively high error due to "approximating a real-life problem, which may be extremely complicated, by a much simpler model" (James et al., 2017, p. 35). In particular, the coefficient estimates in linear regression are likely to be poor estimates of the "true" coefficients in the broader population.

On the other hand, this high bias implies an advantage of linear regression, its relatively low variance. Low variance means that if we were to re-estimate the model on a different training set, our parameter estimates would be relatively stable and change only slightly. Thus, we can be reasonably confident that, if we had randomly selected a different set of observations to comprise our training set, our coefficient estimates and model fit would not substantially change.

Several researchers have used linear regression approaches in other settings to determine determinants of local economic growth. For instance, Rupasingha, Goetz, and Freshwater (2000) look at the effects of social capital, as measured by factors such as membership in voluntary associations, on county-level growth in income per capita over 1990 through 1996. They find that "social capital or civic engagement is an important independent determinant of economic growth in U.S. counties" (p. 571). Similarly, Thiede and Monnat (2016) use linear regression to determine the effect of various county-level characteristics, as measured in 2006, on changes in county unemployment during the Great Recession, from 2006 through 2009. They find that counties with larger minority populations and a higher share of employment in manufacturing tend to have experienced larger increases in unemployment during the Great Recession. Importantly, these studies both consider only one dependent variable. This study considers three dependent variables which all may function as proxies for local economic performance.

Description of Decision Trees

A decision tree uses a tree-like set of if-then statements to group data points into bins. A decision tree in which the target variable is continuous is called a regression tree. In a regression tree, the predicted value of the target variable for a given bin equals the average value of this variable for training set observations in that bin.

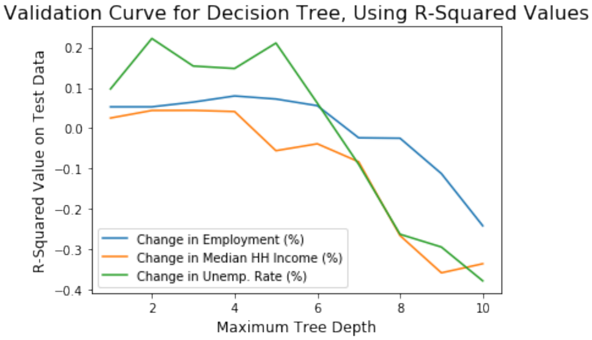
As a simple example, say that we try to predict individuals' earnings based on their years of education and height. A simple decision tree might predict that individuals with at least sixteen years of education have earnings of $100,000; among individuals with fewer than sixteen years of education, those who are at least 70 inches tall have earnings of $70,000, and those below 70 inches have earnings of $60,000.

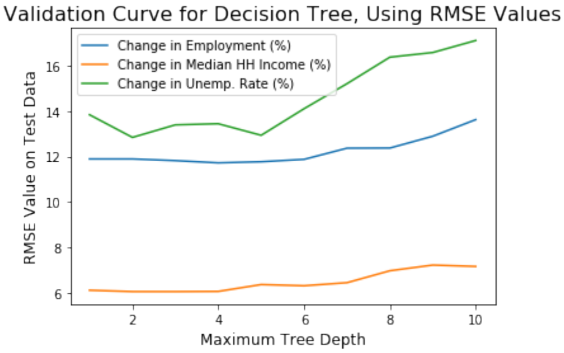
Decision trees are non-parametric, meaning no assumptions of functional form are required. Consequently, decision trees are quite flexible, which may lead to relatively high predictive accuracy and low bias as the tree's increased flexibility allows it to better approximate the real world. Unlike in linear regression, decision trees will automatically discard any unimportant variables, as they will simply not appear in the tree.

Decision trees are slightly less interpretable than linear regression, because there are no parameter estimates that allow us to look at the effect of a one-unit change in a particular explanatory variable. Still, decision trees are easy to visualize, and visualizing a decision tree makes it easy to identify important variables. Additionally, regression trees allow us to calculate the relative variable importances, which indicate each variable's role in reducing the distance between predicted and actual values in the regression tree (Lewinson, 2019). Variable importances are measured as percentages, and for a given regression tree the importances across variables will sum to one. In this paper we use this measure to assess the importance of each variable in the regression tree.

A major disadvantage of a decision tree is that their high flexibility also comes with high variance. That is, decision trees tend to be susceptible to overfitting of data, and a few observations can drastically change the look of the decision tree. Additionally, decision trees can quickly become very complex and large, with many layers (James et al., 2017). An extremely complex tree is no longer very interpretable.

Luckily, one way to partially overcome this limitation is to restrict the decision tree to have a maximum depth (i.e., the maximum number of if-then statements that must be answered before a predicted value can be assigned to an observation). The depth of a decision tree is a hyperparameter, meaning that it must be specified before the model is fit. In order to determine the optimal depth for our decision trees, we loop over numbers from 1 through 10, inclusive, and for each dependent variable fit a decision tree with the specified maximum depth. For each variable, we then choose the maximum tree depth that gives the highest accuracy on the test data, as measured by either R-squared or the RMSE. The plots below show how these two metrics change for each variable as the pre-specified maximum tree depth changes. We can see that R-squared (RMSE) on the test data first tends to slightly increase (decrease) as tree depth increases, indicating improved performance from increased flexibility. However, it then begins to decrease (increase) as the maximum depth approaches ten, a sign that the decision tree is overfitting the training data. Overall, these plots indicate that we should select a maximum depth of 4 when our dependent variable is the change in employment, 3 for changes in median household income, and 2 for the change in the unemployment rate.





While decision trees are a less commonly used technique than linear regression, in recent years they have been increasingly used to answer questions related to economic growth. For instance, Annoni and Rubianes (2016) use a regression tree approach to estimate important determinants of growth rates in GDP across regions in the European Union. They find evidence of clear non-linearity with regards to independent variables, which makes regression trees likely to be more accurate than linear regression. Additionally, a recent report by the Organization for Economic Cooperation and Development has found that an adapted form of a decision tree can be as accurate in predicting countries' GDP growth as more traditional time-series linear regression techniques (Woloszko, 2020).

Analysis

Interpretation of Linear Regression Results

The table below shows the results of our three linear regressions, one for each dependent variable, fit to the training data. As mentioned previously, all variables in the regression are standardized, to make interpretations of coefficients comparable. Each column corresponds to a separate regression, with the dependent variable indicated in the first row of the table. Entries correspond to coefficient estimates, with stars denoting statistical significance. Estimates with at least two stars are significant at the 5% significance level.

	Change in Employment (%)	**Change in Median HH Income (%)**	**Change in Unemp. Rate (%)**
Median Household Income ($)	0.04	-0.39***	0.11*
Unemployment Rate (%)	0.01	0	-0.46***
Poverty Rate (%)	-0.09**	-0.20***	0.19***
Pop. w/ College Degree (%)	0.38***	0.41***	-0.08*
Share of Employment in Manufacturing (%)	0.18***	0.08**	-0.17***
Percent of Votes for Obama in 2008	-0.09***	-0.01	-0.11***
Percent of State's Votes for Obama in 2008	0.02	0.03	-0.05**
State's Tax Burden Rate	-0.18***	-0.01	-0.02

Log(County Population)	1.04***	0.07	0.14
Log(County Employment)	-0.99***	-0.12	-0.06
Log(County Population Density)	0.06	0.02	-0.08*
Constant	0	0	0
Number of Observations	1282	1282	1282
Training R^2	0.23	0.06	0.28
Note: All variables are standardized.			

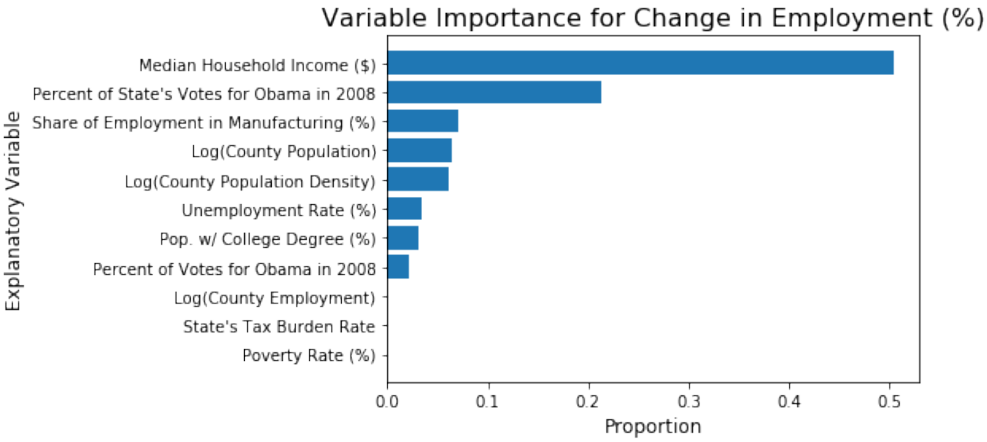
With the change in employment as the dependent variable, we can see that the poverty rate, college-educated rate, manufacturing employment share, county-level votes for Obama, population, employment, and state-level tax burdens are all statistically significant. By far the most important of these variables are county-level population and employment. A one standard-deviation increase in log county population is associated with an increase of 1.04 standard deviations in a county's percentage change in employment, or nearly 11 percentage points. On the other hand, a one standard-deviation increase in log county employment is associated with a **decrease** of 0.96 standard deviations in employment growth, or over 10 percentage points. Given that the correlation of log county population and employment in 2009 is 0.95, this suggests that counties with relatively low employment-to-population ratios in 2009 have had higher employment growth, possibly because these counties lost an especially high number of jobs during the Great Recession. The share of the population with a college degree is also a fairly important variable, with a positive coefficient of 0.38. The other statistically significant variables are relatively unimportant in comparison.

With the change in income as the dependent variable, only median household income in 2009, the poverty rate, the college-educated rate, and the share of manufacturing employment are statistically significant. Of these, median household income and the college-educated rate are by far the most important, with coefficients of -0.39 and 0.41, respectively (corresponding to decreases and increases, roughly, of 2.4 percentage points in income growth). These results indicate that counties that were relatively wealthy in 2009 have enjoyed lower income growth between 2009 and 2016, on average. On the other hand, counties with more college-educated residents have enjoyed higher income growth since the Great Recession.

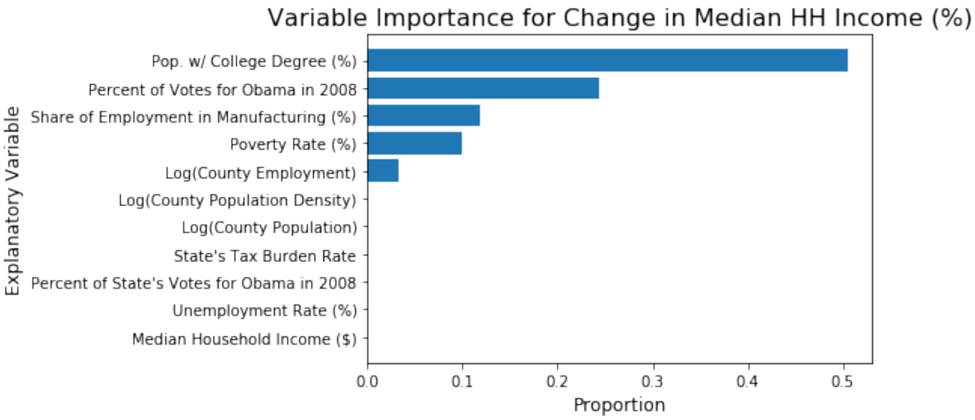
Finally, with the change in the unemployment rate as the dependent variable, the initial 2009 unemployment rate, poverty rate, manufacturing employment share, and county-level and state-level votes for Obama in 2008 are statistically significant. The initial 2009 unemployment rate is by far the most important variable. Its coefficient of -0.46 indicates that a one standard deviation increase in the 2009 unemployment rate is associated with a decrease of 0.46 standard deviations, or 8.3 percentage points, in the percent change in unemployment from 2009-2016. The other significant variables have relatively small coefficients and are comparatively unimportant.

Interpretation of Decision Tree Results

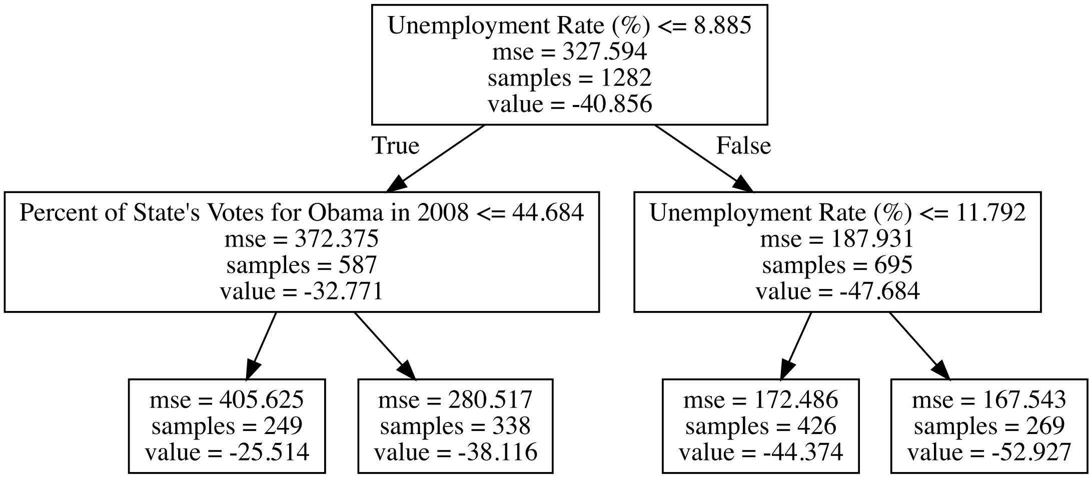
The decision tree for the change in employment has a depth of four. Because of its complexity, we do not visualize the tree here, but we do show the variable importance plot below. We can see that median household income in 2009 is by far the most important variable, accounting for nearly 50 percent of total variable importance. State-level votes for Obama in 2008 are also quite important, but all other variables are of relatively low importance.



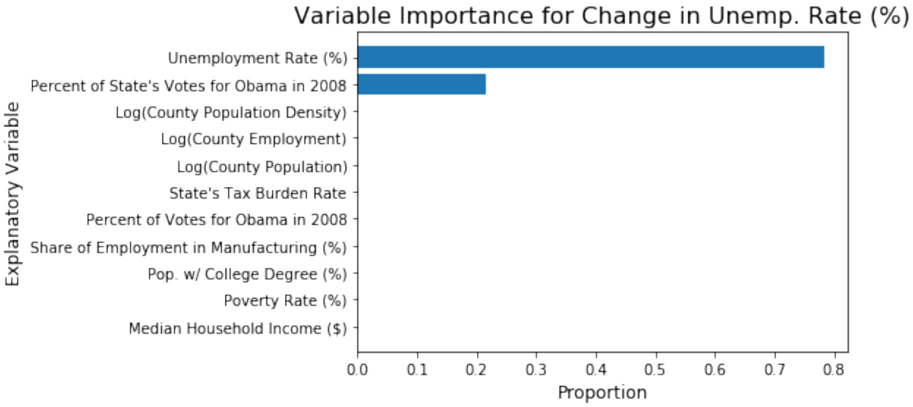
The decision tree for the change in median household income has a depth of three and, because of its complexity, is not visualized here (all decision trees can be viewed as PDF files in the folder accompanying this report). In the variable importance plot below, we can see that the percent of a county's population with a college degree is by far the most important variable, accounting for roughly 50 percent of variable importance. County-level vote shares for Obama in 2008 account for roughly one-quarter of total importance, with the poverty rate and manufacturing employment share each accounting for slightly over ten percent. These four variables thus seem to be by far the most important in predicting median household income using a decision tree.



The decision tree for changes in the unemployment rate has a depth of two and is shown below. On one extreme, counties which had an unemployment rate below 8.885% in 2009 and are in states that gave Obama less than 44.7% of their vote have predicted declines in the unemployment rate of just 25.5%. On the other extreme, counties with 2009 unemployment rates above 11.792% have predicted declines in the unemployment rate of 52.9%. This tree paints a general picture that counties with higher unemployment rates in 2009 have had steeper unemployment declines between 2009 and 2016.



The variable importances plot for changes in the unemployment rate is shown below. As only two variables appear in the decision tree, they account for all of the total variable importance. The initial 2009 unemployment rate is by far the most important variable, accounting for nearly 80% of total variable importance.



Evaluation and Prediction, and Discussion of Limitations

The preceding discussion has focused on inference, particularly determining which variables are most important in explaining our three response variables. Here, we consider the models' predictive accuracy and explanatory power, as measured by R-squared and RMSE (both defined earlier). The table below shows the R-squared and RMSE values from each model applied to the test data. While the RMSE in the linear regression is based on standardized values, in the table below it is multiplied by the standard deviation of the variable, to give a comparable RMSE to the RMSE from the decision tree.

Dependent Variable	**Estimation Method**	**R-squared**	**RMSE**
Change in Employment (%)	Linear Regression	0.136	9.773
	Decision Tree	0.080	11.720
Change in Median HH Income (%)	Linear Regression	0.009	5.890
	Decision Tree	0.044	6.064
Change in Unemp. Rate (%)	Linear Regression	0.287	15.264
	Decision Tree	0.222	12.836

We can see that all of these models have relatively low explanatory and predictive power. The highest R-squared value is just 0.287, for the case of using linear regression to predict changes in the unemployment rate. Therefore, all of these models, including all of our feature variables, are able to explain only a small portion of total variation in our dependent variables of interest. R-squared values for the change in median household income are particularly small, with our linear regression model able to explain just 0.9% of total variation in changes in income. All of the RMSE values are relatively close in value to the equivalent of one standard deviation of their respective dependent variable, which means (roughly) that the average predicted value is close to one full standard deviation away from the corresponding observed value. Overall, these models seem to perform best for predicting changes in the unemployment rate, and worst for predicting changes in median household income.

Comparison of Linear Regression and Decision Tree Models, and Discussion of Limitations

In terms of evaluation and prediction, the results above indicate that both decision tree and regression results tend to have relatively low explanatory and predictive power. Neither analytical approach appears to do systematically better than the other. For instance, when change in employment is the dependent variable, linear regression produces both a higher R-squared and lower RMSE, but when the change in income is the dependent variable, R-squared is nearly five times higher with a decision tree than with linear regression. There does not seem to be a clear justification for favoring one approach over the other.

In terms of determining important explanatory variables, linear regression and decision trees sometimes, but not always, produce fairly similar interpretations. With the change in the unemployment rate as the dependent variable, both methods found that the unemployment rate in 2009 is by far the most important variable. In particular, both models predict that higher unemployment rates in 2009 are associated with steeper declines in unemployment through 2016. Essentially, counties that had the most severe increases in unemployment during the Great Recession tended to have among the strongest recoveries in the unemployment rate.

With the change in total employment as the dependent variable, linear regression found that log population and log employment were by far the most important variables, with the share of the population that is college-educated a somewhat distant third. In contrast, a decision tree found median household income and state-level vote shares for Obama to be by far the most important variables. Log population and the rate of college education both have importances below 10%, with log employment having an importance of 0% (meaning it does not appear at all in the decision tree). Thus, the two models are not in substantial agreement. It is possible that this disagreement may arise in part from the fact that log population and log employment are extremely highly correlated, making it difficult for the models to tease out the effects of one of these variables as opposed to the other.

With the change in median household income as the dependent variable, linear regression found that the college-educated share and median household income in 2009 were by far the two most important variables, with coefficients of opposite sign but very similar magnitudes. A decision tree found that the college-educated share was by far the most important variable, with a variable importance of roughly 50%. However, median household income in 2009 does not appear at all in the decision tree. Additionally, the decision tree gives a roughly 25% variable importance to Obama's county-level vote share, while this variable is statistically insignificant in linear regression. Thus, in this case linear regression and decision trees seem to be in only modest agreement.

Conclusion

Findings and Limitations

Several key findings emerge from this analysis. First, it is worth re-iterating that our three dependent variables were found to be weakly correlated with each other. Because each of these variables is meant as a proxy for economic performance, the weak correlations mean that all three of these variables should be considered and evaluated separately.

Second, both the linear regression and decision tree models considered in this report have relatively low explanatory power for all three variables, and tend to make predictions that are not extremely accurate. This suggests that the set of eleven independent variables considered in this text can only weakly explain variations in county-level economic performance since the Great Recession. Neither of the two methods appears to perform systematically better, or worse, than the other.

Third, the linear regression and decision tree models generally have only a modest degree of agreement when applied to the same dependent variable. A slight exception is with the unemployment rate as the dependent variable, where both models agree that the unemployment rate in 2009 is by far the most important predictor of declines in the unemployment rate.

Fourth, there is no single variable that consistently appears to be highly important in explaining all three dependent variables. That is, there is no key metric that appears to strongly predict changes in employment, the unemployment rate, and median household income. Rather, separate explanatory variables seem to be important in explaining each dependent variable.

While limitations of the data have already been explained in detail, there are several key limitations of the analysis. Namely, the weak explanatory power of the models, combined with the lack of consistent findings across our dependent variables, makes it difficult to draw strong conclusions. No particular variable or factor seems to be an especially strong predictor of all three of our measures of economic performance. Additionally, the limitations of available data meant that local county-level policy characteristics could not be directly considered; it is precisely the effects of these variables that would probably be most of interest to policymakers. Thus, future work should try to incorporate more policy-level variable characteristics, even if at a level of geography other than the county level.

Ethical Considerations

Because this was an observational study relying exclusively on data available from various government surveys, this study does not raise any major ethical issues. However, ethical concerns may arise if policymakers or other actors use the results of this study to make policy decisions. Our research question was interested in determining the factors that most strongly seem to determine county-level economic performance since the Great Recession. Due to a lack of available data at the county level, we were unable to consider any variables that can be directly set by policymakers, with the exception of state-level tax burdens. Thus, we do not believe that the results of this study can be easily used by policymakers to determine local policy, which might shape economic outcomes for residents of these communities, a result that comes with ethical implications for these communities. Given the poor explanatory power of the models considered in this report, we would explicitly caution policymakers against drawing any conclusions regarding how to shape policy to promote economic growth.

Works Cited

All Employees, Total Nonfarm. (2020, February 7). Retrieved February 17, 2020, from https://fred.stlouisfed.org/series/PAYEMS

Amadeo, K. (2019, November 14). 4 Ways to Measure Income per Person. Retrieved April 5, 2020, from https://www.thebalance.com/income-per-capita-calculation-and-u-s-statistics-3305852

Annoni, P. (2016). Tree-based approaches for understanding growth patterns in the European regions. *Region*, *3*(2), 23. doi: 10.18335/region.v3i2.115

Autor, D., Dorn, D., Hanson, G., & Majlesi, K. (2017). Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure. doi: 10.3386/w22637

Bailey, M. A. (2016). *Real econometrics: the right tools to answer important questions*. New York: Oxford University Press.

Blankley, B. (2019, February 27). Report: Blue states have higher property taxes than red states, with New Jersey and Illinois having the highest in U.S. Retrieved April 5, 2020, from https://www.thecentersquare.com/national/report-blue-states-have-higher-property-taxes-than-red-states/article_9c75b456-3a27-11e9-a1c1-9bc3d8e2355f.html

Cassidy, J. (2018, September 13). Ten Years After the Start of the Great Recession, Middle-Class Incomes Are Only Just Catching Up. Retrieved April 5, 2020, from https://www.newyorker.com/news/our-columnists/ten-years-after-the-start-of-the-great-recession-middle-class-incomes-are-only-just-catching-up

Chart Book: Tracking the Post-Great Recession Economy. (2020, April 3). Retrieved April 5, 2020, from https://www.cbpp.org/research/economy/chart-book-tracking-the-post-great-recession-economy

Classen, T. J., & Dunn, R. A. (2012). The effect of job loss and unemployment duration on suicide risk in the United States: a new look using mass-layoffs and unemployment duration. *Health Economics*, *21*(3), 338–350. doi: 10.1002/hec.1719

Cunningham, E. (2018, April). Great Recession, great recovery? Trends from the Current Population Survey : Monthly Labor Review. Retrieved May 4, 2020, from https://www.bls.gov/opub/mlr/2018/article/great-recession-great-recovery.htm

Error Measurement. (n.d.). Retrieved April 5, 2020, from https://www.bls.gov/opub/hom/topic/error-measurements.htm

Florida, R. (2018, August 30). Why Places With Strict Land-Use Restrictions Vote Democratic. Retrieved April 5, 2020, from https://www.citylab.com/life/2018/08/how-land-use-restrictions-make-places-tilt-left/568780/

Hollingsworth, A., Ruhm, C., & Simon, K. (2017). Macroeconomic Conditions and Opioid Abuse. *Journal of Health Economics*, *56*, 222–233. doi: 10.3386/w23192

How many counties are in the United States? (n.d.). Retrieved February 17, 2020, from https://www.usgs.gov/faqs/how-many-counties-are-united-states?qt-news_science_products=0#qt-news_science_products

How the Government Measures Unemployment. (2015, October 8). Retrieved April 5, 2020, from https://www.bls.gov/cps/cps_htgm.htm

Hsieh, C.-T., & Moretti, E. (2019). Housing Constraints and Spatial Misallocation. *American Economic Journal: Macroeconomics*, *11*(2), 1–39. doi: 10.1257/mac.20170388

IPUMS NHGIS. (n.d.). Retrieved April 5, 2020, from https://www.nhgis.org/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R*. New York: Springer.

Lewinson, E. (2019, April 11). Explaining Feature Importance by example of a Random Forest. Retrieved from https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e

Local Area Unemployment Statistics Home Page. (n.d.). Retrieved April 5, 2020, from https://www.bls.gov/lau/

McCann, A. (2019, April 2). Tax Burden by State. Retrieved April 5, 2020, from https://wallethub.com/edu/states-with-highest-lowest-tax-burden/20494/

Margins of error in the ACS. (n.d.). Retrieved February 17, 2020, from https://walkerke.github.io/tidycensus/articles/margins-of-error.html

MIT Election Lab. (2020, April 9). Retrieved May 4, 2020, from https://electionlab.mit.edu/

Real Median Household Income in the United States. (2019, September 10). Retrieved February 17, 2020, from https://fred.stlouisfed.org/series/MEHOINUSA672N#0

Rupasingha, A., Goetz, S. J., & Freshwater, D. (2000). Social Capital and Economic Growth: A County-Level Analysis. *Journal of Agricultural and Applied Economics*, *32*(3), 565–572. doi: 10.1017/s1074070800020654

Sauter, M. B. (2018, October 12). 10 years later, these 28 U.S. cities never recovered from the Great Recession. Retrieved February 17, 2020, from https://www.usatoday.com/story/money/economy/2018/10/12/cities-never-recovered-great-recession/38094325/

Schneider, H. (2019, July 19). As U.S. 'superstar' cities thrive, weaker ones get left behind. Retrieved February 17, 2020, from https://www.reuters.com/article/us-usa-economy-nashville-insight/as-u-s-superstar-cities-thrive-weaker-ones-get-left-behind-idUSKCN1UE13B

School Census Data. (n.d.). Retrieved February 17, 2020, from https://www.greatschools.org/api/docs/school-census-data/

State-Local Tax Burden Rankings. (2016, January 20). Retrieved May 4, 2020, from https://taxfoundation.org/publications/state-local-tax-burden-rankings/

Tax Burden State By State. (n.d.). Retrieved April 5, 2020, from https://www.forbes.com/pictures/emeg45ehhij/tax-burden-state-by-sta/#3c696bbf3a67

The Recession of 2007-2009. (2012, February). Retrieved April 5, 2020, from https://www.bls.gov/spotlight/2012/recession/pdf/recession_bls_spotlight.pdf

Thiede, B., & Monnat, S. (2016). The Great Recession and America's geography of unemployment. *Demographic Research*, *35*, 891–928. doi: 10.4054/demres.2016.35.30

US Census Bureau. (n.d.). CBP Datasets. Retrieved February 17, 2020, from https://www.census.gov/programs-surveys/cbp/data/datasets.html

US Census Bureau. (n.d.). Gazetteer Files. Retrieved from https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html

US Census Bureau. (2019, September 26). American Community Survey Data. Retrieved February 17, 2020, from https://www.census.gov/programs-surveys/acs/data.html

US Census Bureau. (2019, December 12). Methodology. Retrieved February 17, 2020, from https://www.census.gov/programs-surveys/cbp/technical-documentation/methodology.html#par_textimage_245304869

US Census Bureau. (2019, September 17). When to Use 1-year, 3-year, or 5-year Estimates. Retrieved February 17, 2020, from https://www.census.gov/programs-surveys/acs/guidance/estimates.html

Wharton Residential Land Use Regulation Index (WRLURI). (n.d.). Retrieved February 17, 2020, from http://real-faculty.wharton.upenn.edu/gyourko/land-use-survey/

Woloszko, N. (2020). Adaptive Trees: a new approach to economic forecasting. *OECD Economics Department Working Papers*. doi: 10.1787/5569a0aa-en