2016 Presidential Elections

Using demographic and socio-economic factors of the U.S. population, which candidate will prevail on a county by county basis for the states of Ohio and Florida?

URP 4273
Juna Papajorgji, PhD
4/18/2016

Group Members
Austin Bouchard, Christian Breault, Sky Button, Natalie Castaldo

Main Goal

The main objective of our research is to analyze twelve of the same indicators for each county in the states of Florida and Ohio, both swing states, to predict which presidential candidate has the advantage in the upcoming general election.

Background/Problem Statement

Starting in 1789, there have been a total of 57 presidential elections in U.S. History as of 2012. Taking place once every four years, the presidential election and campaigns associated provide an exciting time for America to exercise their right to vote with the freedom of their choice. The elections go through many series of events including the primary elections, leading up to the nomination of a candidate on each party's behalf, and then finally the general election held in November, when the new President will be decided.

Just like presidential elections have been around for years, so have correlation and prediction studies. Forecasting has become popular not just in weather, but in a wide-range of fields. From using correlational data to study the synchronization of a mother and her child, to using prediction models to study bankruptcy, the possibilities of what you can do are endless. Companies even hire big data and analytics firms to identify these sorts of patterns to give their business better opportunities.

Each presidential election has its own distinctive characteristics that make it a little different than the others. From the first president, George Washington, elected with no competition, to our current president, Barack Obama, the first elected African-American president, each election provides a new opportunity to analyze voter tendencies. The upcoming election in November 2016 will also be an interesting one. On one side, Hillary Clinton, representing the Democratic Party, could become first woman president elected into office. On the other side, Donald Trump, representing the Republican Party could become the first president since Eisenhower who has never had any previous government experience, and the first president who was a former reality TV star. Either result would be a compelling one. After the careful selection of influential variables in relation to different demographics and socio-economic conditions, and through the examination of how these variables correlate with how people vote, we hope to come up with an accurate prediction for Ohio and Florida.

Scope and Characteristics of the Study Area

Our study focuses on two states that are infamous for being influential swing states in the United States, and more specifically, the states of Florida and Ohio. Using secondary data on demographics and socio-economic factors, we sought out what we felt were the most important indices to analyze. We chose the following indicators based on areas that have proven a strong connection to either the Republican or Democratic Party. In order to determine which candidate would prevail in each state, we examined specific indicators within the following four main categories: Population, Education, Political, and Financial.
Underneath the Population profile, data was collected from census.gov for the percent of whites, percent of hispanics, percent of foreign born persons, voter age brackets, and the number of males and females within each county for each state. For the Education profile, the percent of high school educated and the percent of college educated persons were collected and analyzed. For the Political profile, data was obtained from the trends of the 2008 presidential election, as well as the 2016 primary election. Lastly, for the Financial profile, data on the income brackets for the lower, upper, and middle classes were obtained as well as the percent of households in each county that make less than $16,000 per year.

Objectives for Accomplishing the Main Goal/ Criteria

In order to effectively project the outcome of the 2016 presidential election, we established a number of key objectives. One important objective was to create a list of demographic features based on recent data (no more than eight years old) that would be useful for determining which candidate should be favored in each county, based on the voting trends displayed by each demographic. Some variables had to be left out to prevent our data from being too one-sided, such as 2012 election results, which would have given Clinton an unfair advantage since we'd already used 2008 election results that favored democrats as one of our indicators.

Once we determined suitable indicators, our next objective was to extract relevant demographic data from credible sources, and cite those sources to ensure we had reliable data that would enable us to accurately analyze the 2016 election. Most of our data sources came from the Census Bureau's publicly distributed data, but a small portion also came from online articles that sourced to the Census Bureau or another respectable distributor.

With proper data and indicators in place, our next objective was to construct a table for both Florida and Ohio that organized the data in an easily sortable format. We also sought to develop input a formula that would determine the amount of hypothetical voters for each candidate. With the formula in place, our next objective was to join the resulting table to an ArcGIS basemap, so that we could view, manipulate, and geospatially symbolize the data in ArcMap.

Our final objective was then to take raw Florida and Ohio basemaps and format them so that they would display each county's projected winner, with red shades on the map favoring Trump and blue shades on the map favoring Hillary. When creating this map, we made color shades lighter to indicate lower margins of victory for either candidate, and darker to indicate stronger, more landslide victories for each candidate.

Methodology

In order to receive accurate results from our experimentation, a certain methodology was appropriated. Our prediction is based upon procuring sets of influential indicator data and predicting results based upon the relative advantage each candidate has in those demographics. The indicators that were chosen were based off of popular indicators that are believed to have a significant impact on the 2016 Presidential Election. These variables include demographics based upon income class, race/ethnicity, past election results, age, and level of education. Each indicator was then grouped into different categories to make up a Population, Education, Political, and Financial Profile as previously mentioned. All variables were converted into points based upon the population density of each county so that all data is represented in the same format.

Once the data was collected, Equation 1 was used for each county in order to determine the candidate that had the greatest likelihood of garnering the most votes:

Eqn 1.

$$\text{Point Total} = \sum (\text{Weighted Indicator Value for Candidate}) * (\text{Indicator Value}) * (\text{Voter Turnout})$$

The weighted indicator value for each candidate is a value from 0-1 that represents how much of an impact each indicator has on the general election. These weighted indicator values were based off of the percentage of votes each demographic represented in past elections and correlation studies.

The values are then multiplied by a Voter Turnout factor that represents how likely each demographic is to go out and vote. Each indicator value is based upon a correlation study that provided statistical analysis on voter turnout. Not all indicators have a Voter Turnout factor associated with it. Figure 1 provides a good example of our equation used in practice.

Figure 1.

For Example,

Indicator= Hispanic Population = 10,000

Weighted Indicator Value for Trump = Percent of Hispanic Population that votes for Trump based off Correlation study = 0.2

Voter Turnout = 70% of Hispanics voted in 2008 Presidential Election = 0.7

Points that Trump gets for Hispanic Population Indicator = 2,000

Note: All values used in this example are theoretical and do not represent actual county data

*All weighted values will be supported by a correlation study done by a reputable source
*Whichever candidate obtains more "points" in a county will be considered the winner of said county.
*Whoever wins the most counties in a state will be considered the winner for that state.
*The equation used is not to be used as a representation of vote totals but rather the differential between vote totals will represent the likelihood that the candidate will win that county/state.

Table 1 provides us the influence and voter turnout factors for each indicator used in our study.

Indicator Name	**Influence Factor (Democrats)***	**Voter Turnout Factor**	**Reference**
Upper Class	0.47	0.75	Link (1) (2)
Middle Class	0.51	0.58	Link (1) (2)
Lower Class	0.63	0.43	Link (1) (2)
Below $16,000	0.72	0.41	Link (1) (2)
At Least High School	0.49	0.50	Link (1) (2)
At Least College	0.52	0.68	Link (1) (2)

Caucasian	0.43	0.6	Link
Foreign Born	0.82	0.45	Link
Hispanic	0.67	0.42	Link
Male	0.52	0.28	Link
Female	0.46	0.28	Link
2008 Presidential Election - Obama	1	1	Link
2008 Presidential Election - McCain	1	1	Link
2016 Primary Election - Hilary	1	1	-
2016 Primary Election - Trump	1	1	-
Ages 20-34	0.59	0.52	Link
Ages 35-44	0.48	0.6	Link
Ages 45-64	0.46	0.7	Link
Ages 65+	0.45	0.7	Link

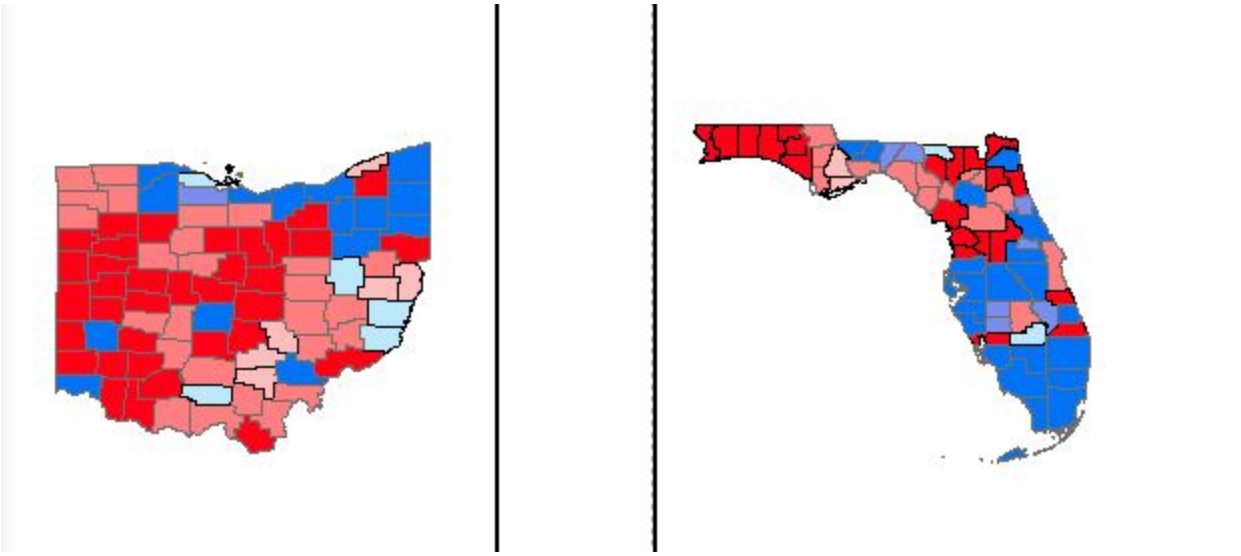
*Repulican influence factor is (1 - Democrat influence factor)

While the data analysis used gives us a potential winner for each county, it is not to say that the winner will positively win that specific county. The data analysis only gives us a potential winner based off of the indicators that are used in the study. Actual election results depend on more than the indicators used and will differentiate greatly to the study that we have produced. However, despite this, our study still gives us a potential based off of the indicators that we deemed important. The larger the differential between the two candidates, the more confident the study is in selecting that candidate as the winner. These confidence intervals are based off the differentials and then scaled based upon a margin of victory. This margin of victory is better explained in the Results and Discussion section below.

Results and Discussion

Our initial analysis resulted an output that was largely biased towards Clinton, who won every county in our simulation. Since this result is nearly impossible, we changed our analysis as described in our methodology. This new approach yielded what was clearly a more reasonable set of maps. The map generated for Ohio was nearly identical to the one from the 2008 election, although Clinton won one more county than Obama. Trump was simulated to win more counties than Clinton in Ohio by a landslide, but Clinton was projected to win counties with higher populations.

Figure 2.



Above: simulated results maps for Florida and Ohio. Blue colors represent Democratic victories, while red colors represent Republican victories. Darker shades of a color indicate a more significant margin of victory.

In Florida, Clinton was projected to do significantly better than Obama did in 2008. Our simulated showed her winning most South Florida counties, and counties containing nearly every major urban center in the state. Despite this, Trump still won more counties than Hillary in Florida, although not by a landslide. Moreover, only a single county (Glades County) that Hillary was projected to win could be considered a "close race" by any standards based on our simulation. We therefore have a high degree of confidence in almost all of Hillary's victories in Florida. The counties won by Trump in the Florida were once again predominantly rural, containing a low number of total voters.

Our model in general was not at all optimistic about Trump's odds in Ohio or Florida, as he was projected to do even worse in both states than McCain did in 2008 -- an election year in which Obama won both Florida and Ohio. Additionally, although our initial output that had Hillary winning every county was clearly biased in her favor, the fact that Trump did not win even a single county in this simulation is worth highlighting, because it demonstrates Trump's potential weakness relative to past Republican frontrunners as this stage in the election cycle. This sentiment has been echoed numerous times recently by statisticians and pundits alike, and makes it obvious why most well-known Republican politicians have not formally backed Donald Trump.

Trump's potential weakness as the Republican party's nominee is once again demonstrated by the crushing victory that Hillary is simulated to have in Florida. Since our Ohio map so closely resembled past election result maps from Ohio, we have reason to believe that our model as a whole was quite accurate, making the fact that Hillary is simulated to win significantly more Florida counties than even Obama did in 2008 quite shocking. Hillary's results in Ohio suggest that she'll win the state in a somewhat close race, in a similar fashion to Obama in 2008. However, the same analysis applied to Florida suggests the Clinton will beat Trump in a landslide victory. Considering that Florida has been considered a swing state in nearly every election, this is extremely unusual, and once again supports the idea that Trump is one of the weakest Republican frontrunners in recent memory.

Despite our results however, it's important to note that intangible and immeasurable factors in Florida could have caused our analysis to be less effective in this state. For example, it's possible that Trump's ability to market himself as a "non-politician" is better in Florida due to complex socio-geographic factors that we weren't able to consider, which might have led us to underestimate his vote totals. However, there is no evidence for any glaring oversights that would have drastically altered our results, especially since our simulation in Ohio so closely matches up with past election results.

While there was evidence to suggest aberrant election results in Florida this election cycle if Clinton and Trump are the nominees, many trends common in other elections held true. For example, Hillary won almost all counties with big cities, such as those containing Cincinnati, Cleveland, Columbus, Tampa, Miami, Orlando, and Tallahassee. Meanwhile, Trump won the majority of the counties despite clearly being simulated to lose the overall majority of the vote in both states. It is typical in most elections that most of the Democratic vote is won in areas with high population densities, while a large portion of the Republican vote comes from areas with lower population densities. This means that ultimately the deciding factor in many elections is how people in suburban areas vote. Our simulation would therefore suggest that Hillary would do better than Trump in most suburban areas, if both candidates were their respective parties nominees. However, there are a few races in Ohio that we projected to be close but favoring Hillary, largely in suburban areas, so it is possible that if Trump managed to win in these areas, he'd have at least some chance of winning Ohio, however small. On the other hand, quite a few

of the victories simulated for Trump in Ohio are also projected to be close races (in which we therefore have a low level of confidence in projecting Trump as the winner), so it's also possible that Trump could do even worse in Ohio than we've projected, which could potentially yield an unusual landslide victory for Hillary in this state as well.

Conclusion

In conclusion, as we see from the results, Trump is simulated to win more counties but Hillary is simulated to win more of the votes overall. Most of Hillary's votes come from densely counties that hold cities such as Miami whereas most of Trump's votes come from much smaller counties. For our simulated maps, Hillary received more counties and votes in Florida than Obama did in 2008 but also received an almost equal amount of counties and votes as Obama's campaign in Ohio in 2008. There are a few counties with a near-equal amount of votes for Trump and Hillary, such as Glades county in Florida, which is very small at approximately 13,000 residents. We assume that this county has a near-tie in votes because it is located near the heavily democratic region of Southern Florida but also has some belief in Trump being a low population county. Overall, we believe our simulation was accurate given the similar trends in our projections relative to the map of the 2008 primary election votes.

References

1. "A Deep Dive Into Party Affiliation." *Pew Research Center for the People and the Press RSS*. People Press, 07 Apr. 2015. Web. 20 Apr. 2016.
2. "Election Center 2008." *CNN*. Cable News Network, 2008. Web. 20 Apr. 2016.
3. "Election Polls -- Vote by Groups, 2008." *Gallup.com*. Gallup, 2008. Web. 20 Apr. 2016.
4. "Exit Polls." - *Election Results 2008*. NY Times, 2008. Web. 20 Apr. 2016.
5. Gewurz, Danielle. "Party Affiliation and Election Polls." *Pew Research Center for the People and the Press RSS*. People Press, 03 Aug. 2012. Web. 20 Apr. 2016.
6. McDonald, Michael. "Voter Demographics." *United States Election Project*. University of Florida, n.d. Web. 20 Apr. 2016.
7. Minnite, Lorainne. "The University Collaborative New Americans Exit Poll Project." *THE UNIVERSITY COLLABORATIVE NEW AMERICANS EXIT POLL PROJECT* (2008): n. pag. NYIC, 2008. Web. 20 Apr. 2016.
8. Thompson, Derek. "Does Your Wage Predict Your Vote." *The Atlantic*. Atlantic Media Company, 5 Nov. 2012. Web. 20 Apr. 2016.
9. "Voter Turnout By Income, 2008 US Presidential Election." *An Equal Say And An Equal Chance For All*. Stacked Deck: How the Dominance of Politics by the Affluent & Business Undermines Economic Mobility in America, 2008. Web. 20 Apr. 2016.