Paper ###-2020

**Modeling of Political Battleground Counties**

Thomas Coleman, Kendall Lipsey, Tanner Miles, and Mukila Rajasekar; University of North Carolina, Wilmington

# Abstract

As America’s political polarization increases, politics are becoming more important in our everyday lives. In 2016, the US experienced an 18% increase in news-consumption from the year prior (“Nielsen Report”, 2017). Daily, we watched political candidates campaigning for the presidency of the United States. For a candidate to maximize the impact and efficiency of campaigning, he/she must choose locations that subject to have a large change in votes for their direction. For example, Trump campaigning in California would be an inefficient waste of resources due to little to no change in votes from election to election for the republican party, but Trump campaigning in North Carolina (as he did, often) is extremely beneficial. North Carolina is a state that tends to swing allowing the politician to potentially win more votes. We will use data science to create a model to predict the total change in votes cast for one party over the other using county-level demographics. Based on our findings, we will predict which counties/areas would be most beneficial for candidates to campaign to secure the most votes. The paper only focuses on the two major political parties, Republican and Democrat. The paper will target the 2016 election due to its engaging nature, but the data will be sourced from 2012 and 2016. We took the changes from 2012 to 2016 of each variable to work with the trends between those two election years.

# Introduction

Politics are becoming an increasingly important part of Americans’ lives. A Gallup poll found that under the 2016 election, Americans following the election “very closely” spiked from 31 to 39 percent compared to the previous election, with Republicans having the highest closely following percentage (Auter, 2016). How were Republicans able to do this in the 2016 election? Age, voter issues, and state of the economy all had roles to play in turning up the heat on the 2016 election, but we believe there was a main ingredient at play here: the strategic selection of campaigning locations.

President Trump, love him or hate him, dominated the news stage and rally arena. Trump was often known for his large rally size and had almost triple the free news coverage of Clinton (Confessere and Yourish, 2016). Arguably a large contribution to his success, he hosted more rallies/campaigns in 13/15 battleground states compared to Clinton (Smith and Kreutz, 2016). There are two problems that we focus on with choosing battleground counties: 1) What factors, and in what areas, should political strategists consider when evaluating how to receive the greatest number of votes per state? 2) Depending on the strategy, where should time, effort, and money be allocated to win the greatest number of votes per state?

We put our focus on solving these issues through two methods: 1) Created a predictive model to predict which counties are subject to have the greatest change in total votes from one election to the next. 2) Using the findings from method 1, we determine the most efficient campaigning areas. The goal of a campaign/rally is to attract the most people, keeping this in mind, we will base this on population and population density.

When political scientists and strategists plan a path to victory for a presidential candidate, the complexity of the electoral college weighs heavily. Whether to focus on landslide victories in far leaning states or leveling out competitive ‘purple’ states relies on how many votes you can win on a state-level. However, state lines do not adequately illustrate the differences and diversity represented within each state. Therefore, zooming in and investigating nuances found on a county level can help identify more concrete patterns.

# MEthods

Although there are many comprehensive resources available, there was no single data source which had all the desired variables for every county in the USA. Data was collected for the years 2012 and 2016 by FIPS code, a unique code issued to each county.

Information was gathered on unemployment and economic composition by labor type (which was condensed to blue and white collar). Investigated races include White, African American, Hispanic, American Indian, and Asian American. Highest educational attainment was collected on high school, associate degree, bachelor's degree, and graduate degree. This was later condensed to bachelor's degree or higher (combining bachelor and graduate degrees) and associate degree or lower (combining associate and high school). Population density and voter turnout data was collected. Religious affiliation information was gathered including Orthodox Jewish, Black Protestant, Mainline Protestant, Catholic, Evangelical, Total Religious, and other religious affiliations. Median Household income change (from 2008 to 2012 and 2012 to 2016) was collected. Lastly, total vote, Republican vote, and Democrat vote counts were included. Only party frontrunners (republican and democrat) votes were considered.

All the datasets were transposed by FIPS code and year, then merged into one dataset. This dataset, as well as all sources, are available in our [SAS Election Project Information](https://github.com/trm8847/themiddleground/blob/master/SASDataLinks) GitHub folder. Many observations were excluded for Alaska as it is the only state which does not use FIPS codes, but rather boroughs and judicial districts. Also, some vote information was missing for a few counties in Arizona in 2016.

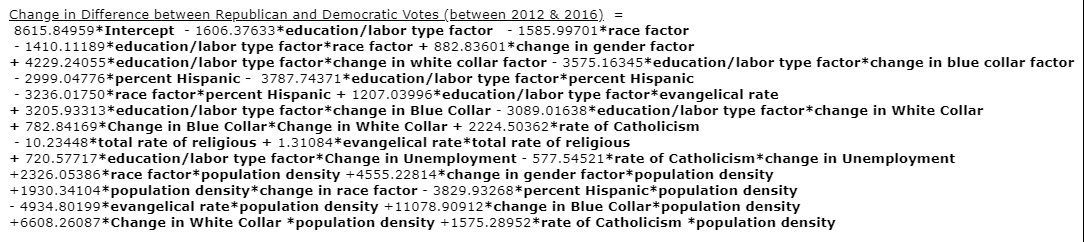
PROC FACTOR was used for dimension reduction from 17 variables to 8 factors. Both the static number represented by each variable (for example, the percentage of the population which is Hispanic) and the change over time (for example, the change in the percentage of the population which is Hispanic) were evaluated. To standardize the variables not used in factors, PROC STANDARD was employed. A response variable was selected to incorporate change over time and party selection. The difference in the number of votes was calculated by subtracting the number of Democrat votes cast (per county, per year) from the number of Republican votes cast. Therefore, a positive number represents the number of votes Republicans received over the Democrat votes cast. The difference was then calculated between 2016 and 2012. Therefore, a negative forty (-40) represents that in 2016 forty more votes went democratic than republican than in 2012. Finally, PROC GLMSELECT was used to model all applicable variables, factors, and their interactions. Our predictive model was produced using PROC GLM. Visualizations were further produced in SAS studio (Figures 1-29) and stored on Github (link following Conclusion).

To properly utilize this predictive model, we included a predicted, p, function and a residual, r, function in our OUTPUT statement of our PROC GLM to find the predicted values and the corresponding residuals for all counties. These values use our model to predict the change in the number of votes for each county from 2012 to 2016. Our residual values help us determine the accuracy of each prediction by comparing our predicted values to our actual values gathered.

# Results

The model resulted in an R-Squared of 0.334595, implying 33% of the variance of the response variable can be accurately explained by the model. The model resulted in a p-value less than .0001, suggesting it is highly statistically significant.

Moreover, when the model was applied to all counties, 419 counties fell within one confidence interval of the model’s predictions.



Model 1. PROC GLM: Modeling Political Votes

The model above illustrates the variables, factors, and interactions.

# Discussion

Due to the numerous variables, interactions, and the multifaceted application of several of the variables, consideration is required when interpreting the model.

## Population Density

One variable worth further investigation is population density. Intuitively, the greater the population density, the more impact other factors have. However, a more nuanced approach shows that, when ordering the variables in terms of the greatest change in votes, interactions with population density rank as the top four. Figure 1 illustrates as the population density increases, the average number of republican votes over democrat votes decreases substantially. However, Figure 2 illustrates the sum (rather than the average) number of republican votes over democrat votes is far greater in less dense areas. Figure 3 highlights the inverse of this relationship. Although there are more votes gathered for democrats in areas with high population density, areas with low population density far outnumber the dense areas with the sum of the votes leaning republican. Of course, as population density is an interaction in other variables in the model, it cannot be considered in isolation. With that being said, the urban-rural divide still proves to have a heavy weight in the model predictions.

## Labor Type

The top two interactions are between population density and the change in labor composition (blue-collar and white-collar). For example, the highest ranked term showing a change of 11,079 votes from the interaction of blue-collar and population density can be interpreted as a hyper actualization of the urban-rural divide paired with labor composition. This shows that if the change of blue-collar workers in a county is one standard deviation below the mean and population density is two standard deviations above the mean (a city), the model predicts a decrease of 22,000 votes which republicans receive above that of democrats.

Figure 4 shows as the change in the percentage of white-collar laborers increases, the average number of republican votes over democrat votes decreases. This relationship can further be seen in the model’s second ranked term (in order of greatest change in votes) is 6,608 change in white collar\*population density. In this instance, a county with a population density one standard deviation above the mean and a change in white collar laborers one standard deviation above the mean, the model would predict an increase of 6608 votes (increase in the number of votes which republicans receive above that of democrats).

Figure 5 shows as the change in the percentage of blue-collar laborers increases, the average number of republican votes over democrat votes decreases. However, further investigation (illustrated in Figure 6) highlights the interaction of population density on labor composition. Figure 6 shows the sum of blue-collar increases leading to a large increase in republican votes, yet a large increase in blue-collar composition is correlated with negligible increase in votes. It is hypothesized that this skewness is indicative of population density as counties with high blue-collar compositions are often of low population density.

## Race

Race also affects the model prediction. Figure 8 shows how racial composition affects the average number of republican votes over democrat votes. Figures 9 and 10 show as the percentage of Hispanic population increases, the average number of republican votes over democrat votes decreases. However, the model’s sixth ranked term (in order of greatest change in votes) is –3,830 percent Hispanic interacting with population density showing that population density and racial composition have a strong weight together. For example, the model predicts an increase of 7,660 republican votes over democrat votes in a county with a Hispanic population two standard deviations less than the mean and population density one standard deviation below the mean. This relationship can be seen in Figure 11. A similar relationship is seen with the African American percentage of a population, represented by the race factor in the model (positive values in this factor indicate a higher percentage of African American population) and can be seen in Figures 13, 14, and 15.

In addition to the racial composition, the change in racial percentages between 2012 and 2016 affect the model predictions. This is best illustrated by the change in White percentage (as it has the largest percentage in most counties). Figure 16 shows that as the change in the percentage of a white population decreases, the average number of republican votes over democrat votes also decreases. The model incorporates this in the form of the change in race factor, where a negative factor represents a higher percentage of white (vs. black) in a population. The model would predict a county with a white population one standard deviation above the mean and a population density one standard deviation below the mean would have 1,930 more republican votes over democrat votes.

## Religion

Religion carries some weight on the model as well. Figure 17 shows a moderate evangelical population (40-45%) is correlated with the highest average number of republican votes over democrat votes. A couple outliers can be noted in counties with 51-100 (per 1000) evangelical participants as these counties include Santa Barbara and San Francisco which lean heavily democrat. Additionally, counties with 250-300 (per 1000) evangelical participants include Houston and Dallas, Texas. The model incorporates the evangelical population in many terms, most notably as an interaction with population density. The model would predict that a county with a population density one standard deviation below the mean and an evangelical population one standard deviation above the mean to receive 4,934 more republican votes over democrat votes.

Figure 18 shows that a moderate catholic population (50-60%) is correlated with the highest average number of republican votes over democrat votes. Outlier counties encompassing Los Angeles (with catholic population of 450-500 per 1000) and in New York and Maine (with catholic populations of 350-400) lean those brackets democrat. The model predicts that a county with catholic population one standard deviation above the mean would result in 2,225 more republican votes over democrat votes. Additionally, the model incorporates Catholicism as an interaction with population density.

The interaction between population density and both catholic and evangelical rates can be illustrated when comparing Figures 20 and 21 (in Figure 22). Although there are more votes gathered for democrat candidates in areas with low religion, areas with low population density and high religion far outnumber the dense areas and the sum of the votes lean republican.

## Geography

Figure 23 displays our best 10% of predictions geographically on a county level. The model is most accurate in predicting counties in the south and spread out through the mid-west. Our model is able to predict 360 counties within 500 votes cast. We believe since population density has shown such strong weight in our model, we see the more rural areas in the south and the mid-west to be more accurately predicted.

Inversely, when looking at figure 24, we see the areas where our model struggled to accurately predict the change in votes to be much denser, higher populated areas. Places such as Los Angeles, New York City, and Miami, all had poor predictions. We believe this to also be due to the strong weight of our population density variable. Since the population density of these areas are far greater than rural areas, this causes our model to severely skew the predictions. This poses an error with our model that we will continue to try and identify and improve on.

Figure 25 shows the difference in total votes between the years 2012 and 2016. As mentioned earlier in the 'Methods' section, positive values or the red zones are associated with Republican votes received and negative or blue zones represent votes received by Democrats. That is, the larger positive values indicate that a large number of votes went to the Republican party in 2016 than in 2012. Similarly, the larger negative values are associated with a large number of votes for the Democratic Party. These big vote difference regions are clearly viewed in Figure 26. For example, in Figure 27 many counties in California have a high concentration of Democratic votes in 2016. The New York state map in Figure 28 shows a high concentration of Democratic votes only in and around the New York City while most of the other counties in this state have higher number of Republican votes.

In addition to these, Figure 29 shows the average residual value for the states as a whole instead of individual counties to give a general idea of the predicted vote change values in our model.

# Conclusion

Through our model we have found some take-aways: race, gender, religion, and labor type affect the number of votes for a party for a given county. Add in population density, and everything changes. The findings of this model, and the highest-ranking variables, should be considered when planning a political campaign.

One issue with our model being used in the future is the changing composition of political parties in the US. Americans are becoming more diverse, educated, and agnostic. For example, as America has become more educated, the education composition of each political party has flipped (Johnson, 2016). With the 2020 election approaching, there is an opportunity for additional data to make our model more accurately portray the current political makeup of the US. In addition, we plan to incorporate updated population and population density, using the 2020 census, to create a model that identifies the most efficient campaigning locations for candidates in battleground areas.

# Project information website

https://github.com/trm8847/themiddleground

# References

Auter, Z. (2019, May 24). Number of Americans Closely Following Politics Spikes. Retrieved January 5, 2020, from <https://news.gallup.com/poll/195749/number-americans-closely-following-politics-spikes.aspx>.

Confessore, N., & Yourish, K. (2016, March 15). $2 Billion Worth of Free Media for Donald Trump. Retrieved January 5, 2020, from <https://www.nytimes.com/2016/03/16/upshot/measuring-donald-trumps-mammoth-advantage-in-free-media.html>.

Johnson, B. (2019, December 31). The changing composition of the U.S. political parties. Retrieved January 5, 2020, from <https://www.people-press.org/2016/09/13/1-the-changing-composition-of-the-political-parties/>.

The Nielsen Total Audience Report: Q4 2016. (2017, March 4). Retrieved January 5, 2020, from <https://www.nielsen.com/us/en/insights/report/2017/the-nielsen-total-audience-report-q4-2016/>.

Smith, C., & Kreutz, L. (2016, November 7). Hillary Clinton's and Donald Trump's Campaigns by the Numbers. Retrieved January 5, 2020, from <https://abcnews.go.com/Politics/hillary-clinton-donald-trumps-campaigns-numbers/story?id=43356783>.

# Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

|  |  |
| --- | --- |
| Thomas Coleman  University of North Carolina, Wilmington  tmc2564@uncw.edu | Tanner Miles  University of North Carolina, Wilmington  trm8847@uncw.edu |
| Kendall Lipsey  University of North Carolina, Wilmington  kdl8569@uncw.edu | Mukila Rajasekar  University of North Carolina, Wilmington  mr4060@uncw.edu |

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.