KNN Classification for Top Demographic Indicators of Voter Shift Between 2016 and 2020 Presidential Elections

Mathematical Evolutions

Andrew Y. Wang

Allison Hahn Elowson, Kristen Fye

Summer Ventures in Science and Mathematics

The University of North Carolina at Charlotte

Abstract

Political analysts have long studied demographic shifts through elections, and with the results of the 2020 presidential election this opportunity comes again. This paper determines the demographic indicators for voter shift between the 2016 and 2020 presidential elections. The K-nearest neighbors classification algorithm is used based on grouped demographic data to return the best grouping, and Principle Component Analysis is used to determine the best variable in the group. Results of the KNN classification showed that sector was the best indicator, with an accuracy of 67.3% across 893 counties. Notably, race was also the poorest predictor, with only a 49% accuracy. This raises interesting questions about the limitations of modern race modelling for voter shift. Principle Component Analysis was used to determine that the professional sector was responsible for the most variance in the employment sector, making the professional sector the best indicator of voter shift.

Keywords: Demographics, voter shift, R, KNN, computational politics

Introduction

The United States electorate has gone through massive demographic shifts in the past decades. It is vital to understand which demographic factors are indicators of voter shift so we can understand the problems facing various groups in today's society. An understanding of these demographic indicators can yield better policy proposals and better voter outreach to increase turnout. Not only was the presidential election of 2020 remarkable for its massive amounts of mail-in voting, but it was also record-breaking for the number of votes cast, the highest in over a century (Rabinowitz, 2020). Electorate shift has been subject to much research due to its insights on the U.S. population and its communities. The 2020 election gives another opportunity for this analysis.

In past election cycles, the role of minority voters has become more pronounced, as it increased the diversity in beliefs of the voting base. Minority groups have commonly tended to favor Democratic candidates due to the anti-immigration policies proposed by many Republican candidates (Daniller, 2020). Since 2000, the Democratic vote percentage among minorities has risen, Democratic candidates winning a majority of votes (Hudak & Stenglein, 2016). Democrats won 65% of the Latino vote in the presidential election of 2000, rising to 73% by 2012 (Hudak & Stenglein, 2016). A similar trend was seen for Asian Americans, going from 57% Democratic in 2000 to 73% in 2012, and African Americans, going from 90% Democratic in 2000 to 93% in 2012 (Hudak & Stenglein, 2016). Based on these trends, as the portion that minority populations make up in the electorate increases, votes for Democratic candidates are expected to do the same.

The urban and rural divide has been another expanding component in voter shift. Urban areas have leaned Democratic, while rural areas lean Republican. In 2018, polling showed 64% of voters in urban areas supported Democrats, while only 38% did in rural areas (amp, 2019). This difference is widening as well. In 2012, Democrats had only a 5% margin over Republicans in urban areas, but this had increased to 17% by 2018 (amp, 2019). This polarization is reflected in rural areas too, the Republican victory margin growing from 29% to 38% between 2012 and 2018 (amp, 2019). Political pundits hypothesize that increased urban will make traditionally Republican states in the Midwest more contested (Beyer, 2016).

While much of past research surrounding demographics and voter shift has been conducted on specific variables, such as race, this study looks at demographic factors holistically, analyzing over 30 factors at once. This study aims to identify the demographic factor or group that best indicates the turnout shift between the 2016 and 2020 presidential elections. The primary contribution of this study is to identify the top indicator of voter shift. A secondary contribution of this study is to analyze the voter shift over the last four years.

Methods

Dataset(s)

The primary dataset used for this analysis was drawn from Kaggle, called "Election, COVID, and Demographic Data by County" and was obtained from https://www.kaggle.com/etsc9287/2020-general-election-polls (Schacht, 2020). This dataset combines the 2017 American Community Survey 5-year estimate, voting results from the 2016 presidential election, and voting results from the 2020 presidential election on a county scale. The 2017 ACS 5-year estimate is a dataset that contains a variety of demographic data on a county scale (Bureau, 2017; Neutrino, 2019). Before any alteration, the dataset consisted of 4,868 rows and 51 columns. Columns 1 - 3 give a number, a name, and a state, in the form of its two-letter abbreviation, to each county. Next, columns 4 - 8 describe the results of the 2016 election, giving the percentage of votes each candidate received, total votes, and votes for each candidate (as a value instead of a percentage). Columns 9 - 13 describe the same things, but for the 2020 election. Following this, columns 14 and 15 give the latitude and longitude of the county. All other columns (16 - 51) describe the demographic factors of the county. While most of the dataset was complete, there was incomplete data towards the end due to Alaska's inconsistent county naming. Table 1 shows the data structure, broken down by the significant column and row groups.

Table 1

Structure of the data

Rows	Type of data	2016 Election	2020 Election	Demographics
1 - 3110	County	Yes	Yes	Yes
3111 - 4658	Extra Districts	No	Yes	No
4659 - 4689	2020 Alaska election data	No	Yes	No
4696 - 4838	Unassigned values	No	Yes	No
4839 - 4868	Alaska demographic data	No	No	Yes

The dataset contains about 1500 rows of extra districts. As extra districts were not U.S. counties, they were missing all demographic data. The ACS only records data for U.S. counties. Therefore, the extra districts were removed from the dataset. This removal was generalized to omit all rows with missing values. However, the data for Alaska was spread out over hundreds of rows due to naming discrepancies at the collection times. Data for the 2016 election results in Alaskan counties was manually added because it was not in the original data. (Elections.alaska.gov, 2017; Thecinyc, 2016). The data for Alaskan counties were merged, and missing data was purged. After the removal of rows with missing data, the dataset went from 4,868 rows to 3,039 rows. The cleaned data accounts for 3,039 of 3,110 total counties.

Finally, packages were imported into the model. This model was constructed in the programming language R using the RStudio development environment. The main packages used with the model were maps, usmap, gridExtra, and tidyverse. Interested readers can find all code used in constructing the model and any figures in the paper at the end of the article.

Data Preparation and Modeling

Voter shift here is defined as the change in margins between the 2016 and 2020 presidential election results. Margins first had to be calculated based on voting data for both presidential elections to analyze the voter shift. Four new columns were added to the data to represent margins and shifts: "margin2016", "margin2020", "shift", and "outcome". Columns "margin2016" and "margin2020" are calculations of winning margin in each county by-election, with "outcome" being a binary indication of shift.

The margin columns do not follow traditional methods of margin computation, as it is necessary for the margin to indicate the victorious party. As opposed to taking the absolute value of the difference, as is the traditional method, the margin was calculated as the difference between the percentage of Democratic votes and percentage of Republican votes (Ballotpedia, n.d.). The margins are calculated as $Margin = Democratic - Republican$. The values range from 1 to -1, where a positive number indicates a Democratic win and a negative number indicates a Republican win. The closer the value to 1 or -1, the larger the victory. The "shift" column was then calculated off of the two margin columns, indicating the strengthening or weakening of a party's margin in a county, calculated by $shift = margin2020 - margin2016$. The values in the shift column range between 2 and -2, with positive values representing a Democratic shift and negative values representing a Republican shift. Finally, the outcome column was added for the KNN model as a binary indication of the direction of the shift, 1 being a Democratic shift and -1 being a Republican shift. One drawback of the shift column is that it cannot portray whether a county has switched parties unless $|shift|$ is greater than 1 (which is not a very likely case). Table 2 shows the added columns and their meanings.

A K-nearest neighbors (KNN) algorithm was applied to the demographic data to analyze the demographic shifts. The KNN algorithm is a machine learning model that can be used as a classification and regression model, basing its learning on K number of clusters. In this study, the KNN algorithm was used as a classification model to predict the

Table 2

Additional columns and their meanings

Added Column	Value Range	Meaning
margin2016	1 to -1	Positive: Democratic win Negative: Republican win
margin2020	1 to -1	Positive: Democratic win Negative: Republican win
shift	2 to -2	Positive: Area got more Democratic Negative: Area got more Republican
outcome	1 or -1	1: Democratic shift -1: Republican shift

margin shift based on demographic data through 3 clusters. Data were split into a group for learning and a group for testing. The KNN algorithm uses the learning group to identify trends in the data and then applies these trends to the testing group. Therefore, demographic variables with a higher percentage of accurate predictions would provide the best indicators of voter shift. As there are 36 columns of demographic data, they were into demographic groups for more straightforward analysis. Each demographic group was composed of various columns. The list below shows how the groups were made.

Population: Columns 16 - 20 and 27. This dataset includes data on COVID-19 cases and deaths, total population, number of men, women, and voting-age citizens. Data here is related to the size of the population.

Race: Columns 21 - 26. This dataset includes data on ethnic backgrounds of people, measured as Hispanic, White, Black, Native (American), Asian, and Pacific (Islander). Data here is related to the racial composition of the county. Originally, data here were based on a count, so the data for race was standardized into a percentage to remove the influence of the total population.

Income: Columns 28 - 30 and columns 31 - 33. This group includes data on income,

income per capita, poverty, and child poverty. Column 30 was removed from the group because it measured error. Data here is related to the income of people in various counties, with poverty as having an income under the poverty threshold.

Sector: Columns 34 - 38 and 44. Columns 34 - 38 describe the sector employment of people in the county as a percentage, giving the percentages for professional, service (service industry), office (office workers), construction, and production. Column 44 gives the percentage of people who work at home. Data in this group describes the employment sector of people in the county.

Transportation: Columns 39 - 43 and 45. Columns 39 - 43 describe the various forms of transportation used by citizens: driving, carpooling, public transportation, walking, or other. Column 45 gives the mean commute in minutes. Data here describes the modes of transportation people use to get to work.

Work: Columns 47 - 50. The values are separated into private work, public work, self-employment, and family work. Data here describes the type of work that people in the county do as a percentage.

The usmaps package was used extensively to create maps of various scales and sizes (Lorenzo, 2020). However, the existing election/demographic data had to be combined with a county dataset from the usmaps package to plot data correctly. This was because the usmaps plot_usmap function required reference to Federal Information Processing Standards (FIPS) values for counties to plot. Combining the two datasets was more difficult than initially anticipated. County names from the usmaps county dataset often had an identifying suffix, such as “county”, “parish”, or “borough”, at the end, and would not directly merge with the demographic and election data. This meant that the county data containing the FIPS codes had to be cleaned of all suffixes before being matched with the existing data. The result was a new dataset of the mutual counties, in which each county now has a corresponding FIPS code for plotting. This new dataset had 3031 rows. With this new data, plots could be made to represent all counties in the country. This dataset could also be scaled down to see certain states or regions of the country.

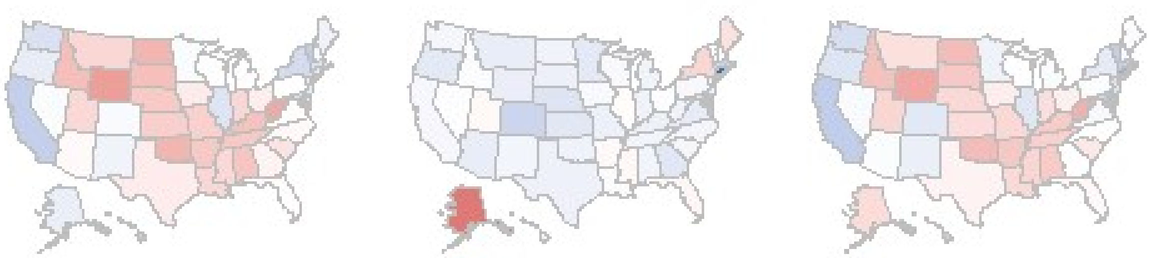
Results

A major component in understanding demographic indicators of voter shift is understanding what voter shift there was. Using the "margin2020", "shift", and "margin2016" columns, the below figure was produced. The data was on a county scale, but because there are over 3,000 counties, the data was averaged through weighted averages to produce a state-level map.

To see the shifts in voter preference between the 2016 and 2020 presidential elections, I used the appended columns of "margin2020", "shift", and "margin2016". The margin and shift data are grouped by state and then put into a map plot, reflected in Figure 1. The darker the shade of blue, the more Democratic a state is. The darker the shade of red, the more Republican a state is.

Figure 1

2016 vote margins, voter shift, and 2020 vote margins at a state scale



Based on the plotting for voter shift, it is evident that most states had a Democratic shift between the 2020 and 2016 presidential elections. While most states shifted Democratic between the two previous presidential elections, notable exceptions to this statement include states like New York, Alaska, and Maine. States with the most prominent Democratic swing were Connecticut and Rhode Island, both with swings of more than 30% in favor of the Democrats. However, most states which shifted Democratic had shifts of between 6-3%. Out of 50 states, only ten states had a Republican shift, with most of these shifts being around 1-2%. Table 3 shows the results of the KNN analysis run

Figure 3

Results of KNN on various demographic groupings

Group	1 : 1	-1 : -1	-1 : 1	1 : -1	Accuracy
Population	232	340	137	184	64.1%
Race	**173**	**264**	**213**	**243**	**49%**
Income	232	296	181	184	59.1%
Sector	**261**	**340**	**137**	**155**	**67.3%**
Transportation	235	275	202	181	57.1%
Work	207	254	223	209	51.6%

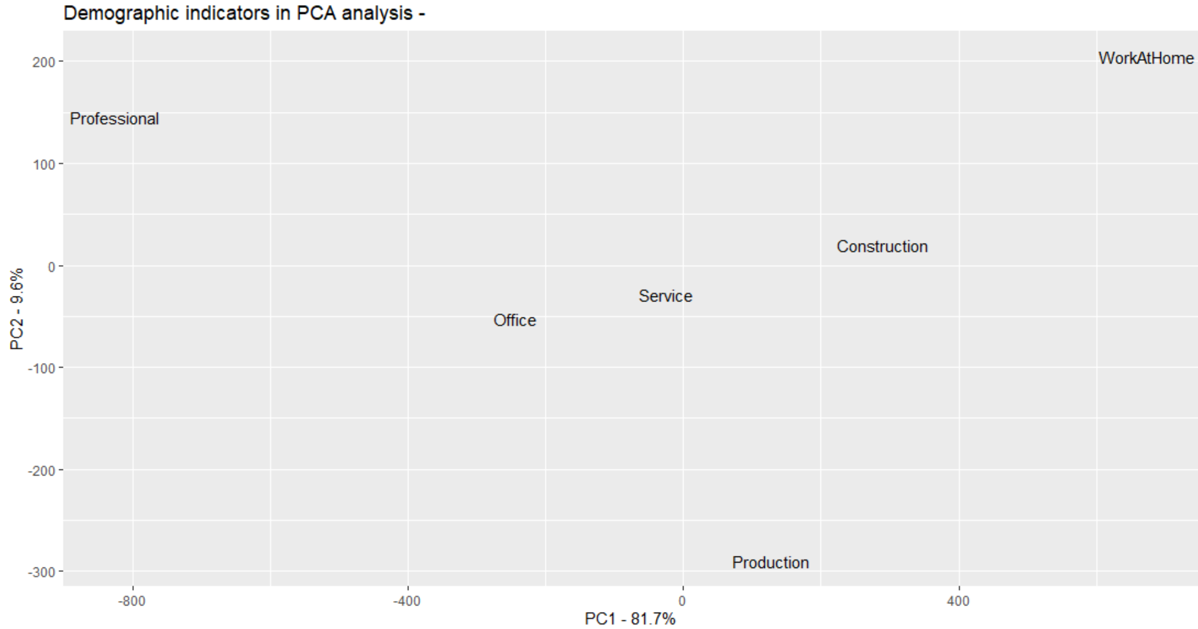
on the six aforementioned demographic groupings. The column headings indicate the four possible scenarios of the KNN model. 1:1 indicates a correct prediction of the Democratic shift, and -1:-1 indicates an accurate forecast of the Republican shift. -1:1 indicates a false Democratic (which the model predicts a Democratic shift but there is a Republican shift), and 1:-1 shows a false Republican (which the model predicts a Republican shift but there is a Democratic shift). The KNN model's results prove that the employment sector is the best indicator of voter shift in the last presidential election, and race is the worst indicator. As the demographic group of the employment sector comprises multiple variables, PCA analysis was then conducted to determine which variable was the most important. The graph of the PCA analysis is reflected in Figure 2, which indicates that the percentage of workers in the professional sector was inversely responsible for the variation in the data.

Discussion

As shown in Figure 1, the United States experienced a largely Democratic shift between the last two presidential elections. The average shift between the two presidential elections on a state scale was 3%. While this statistic disregards the specific population of each state (meaning all states have the same weight), it gives a general idea of voters' opinions across the United States. This Democratic shift between presidential elections also

Figure 2

Results of KNN on various demographic groupings



resulted in five states moving from a Republican win in 2016 to a Democratic victory in 2020. These five states were Arizona, Wisconsin, Michigan, Georgia, and Pennsylvania. While the Midwest remains strongly Republican, the flipping of these five states allowed presidential candidate Joe Biden to obtain the needed 270 electoral college votes to become the president-elect. Furthermore, Democrats also made inroads into Southern states, such as North Carolina, Georgia, and Florida, as the flipping of Georgia marks a significant milestone for Democratic progress in the South. Based on this data, it seems that many previously strong Republican states may become more competitive for Democrats, particularly as areas urbanize. One outlier here is New York, which saw a Republican shift even though it is a traditionally democratic state, and highly urbanized.

Based on Table 3, it is clear that the employment sector is the best indicator for voter shift. This is possibly due to the voter base of the Republican party, which is primarily non-college-educated white people. Voters who make up the Republican party's voting base are more likely to hold jobs in construction, production, or service sectors. Given the widespread unemployment caused by the COVID-19 pandemic, particularly in industries without a heavy digital aspect, it's reasonable to conclude that COVID-19 hit republican voters harder with these economic consequences. This unemployment could then lead to dissatisfaction with the administration and the switching of parties. This claim is further supported by the PCA analysis in Figure 2, which shows that the percentage of professional workers is inverse to variation in the data. This is a fairly straightforward conclusion. The professional sector is usually the majority. All columns add up to 100%; as the majority sector decreases, the other sectors increase, hence its inverse relationship to variation. More surprisingly, however, is the inaccuracy of race as a predictor of shift. With an accuracy of 49%, you could achieve better accuracy by flipping a coin. This inaccuracy is a fascinating result because race has traditionally been one of the top demographic factors for political pundits to model election turnout. Yet, its poor accuracy may call into question some of our assumptions about voting patterns based on ethnicity. Race may be a poor indicator because of the nuances it misses in logging data. Not all ethnic groups share the same geographic trends, so the generalization misses out on nuances and becomes inaccurate. For example, Puerto Ricans traditionally vote Democratic, while Cubans traditionally vote Republican. Yet, they are grouped under the same umbrella of Latino.

Conclusion

This paper found that the best demographic indicator of voter shift between the 2016 and 2020 U.S. presidential elections was the employment sector. Basic arithmetic and plotting were used to visualize areas of significant change to find out voter shift. These calculations showed that most states became more Democratic over the last four years. This conclusion was also reflected in the 2020 presidential election, as five states flipped from Republican (in 2016) to Democratic (in 2020). A combination of KNN classification and PCA were used to derive the results.

Some limitations remain to be addressed in future research. Around 100 counties were missing from analysis due primarily to naming discrepancies. Particularly in Alaska and

Vermont, analysis in those locations may not have been accurate due to these discrepancies.

This research raises interesting questions regarding the future of voting trends in the United States and the current methods we use to predict turnout. Race may be weaker than previously thought, mainly due to its overlooking diversity among our current groupings. The application of electorate demographics to voter outreach and engagement will likely become more salient in the future, as an understanding of what certain voter groups care about will allow for more targeted appeals.

Availability of data and materials

The data for this work was obtained from https://www.kaggle.com/etsc9287/2020-general-election-polls. It was also obtained in part from the R usmaps package.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

This paper is solely the work of the author. All references are included in the bibliography and are cited appropriately.

Acknowledgements

The author appreciates the efforts of the American Community Survey, the U.S. Census Bureau, and FiveThirtyEight in the collection of data and of Kaggle user Ethan Schacht for creation of the dataset used.

References

amp, U. N. (2019). *How the rural vs. urban mindset fuels today's politics.* www.usnews.com/news/best-states/articles/2019-05-14/demographic-shifts-incities-and-states-bring-political-changes-too

Ballotpedia. (n.d.). *Margin-of-victory (mov).* Ballotpedia.org,%20Ballotpedia,%20ballotpedia.org/Margin-of-victory(MOV)

Beyer, S. (2016). *The republican party's urban problem.* www.forbes.com/sites/scottbeyer/2016/05/17/the-gops-urban-problem/?sh=7572b2244662

Bureau, U. S. C. (2017). *Selected economic characteristics.* data.census.gov/cedsci/table?tid=ACSDP5Y2017.DP03

Daniller, A. (2020). *Americans' immigration policy priorities.* www.pewresearch.org/fact-tank/2019/11/12/americans-immigration-policy-priorities-divisionsbetween-and-within-the-two-parties/

Elections.alaska.gov. (2017). *State of alaska - division of elections voters history by age report 16genr-2016 general election.* www.elections.alaska.gov/election/2016/General/VoterHistoryByAgeReport.pdf

Hudak, J., & Stenglein, C. (2016). *How demographic changes are transforming u.s. elections.* www.brookings.edu/blog/fixgov/2016/09/13/how-demographic-changes-are-transforming-u-s-elections/

Lorenzo, P. D. (2020). *2. mapping the us.* cran.r-project.org/web/packages/usmap/vignettes/mapping.html

Neutrino, M. (2019). *Us census demographic data.* www.kaggle.com/muonneutrino/us-census-demographic-data

Rabinowitz, K. (2020). *2020 turnout is the highest in over a century.* www.washingtonpost.com/graphics/2020/elections/voter-turnout/

Schacht, E. (2020). *Election, covid, and demographic data by county.* www.kaggle.com/etsc9287/2020-general-election-polls

Thecinyc. (2016). *Alaska presidential results by county, 1960-2016: Maps.* www.thecinyc.com/alaska-pres-results-by-county-equiv

Additional Files

R Code for this work

The R code used in this analysis can be found at its GitHub repository https://github.com/aywang71/2020-demographic-indicator-ml, with accompanying cleaned data.