The Divided (But Not More Predictable) Electorate: A Machine Learning Analysis of Voting in American Presidential Elections

Seo-young Silvia Kim*† and Jan Zilinsky‡

†American University

‡New York University

March 16, 2021

Abstract

Partisan sorting by social groups is believed to increase affective polarization and decrease group-level leverage in representation. Mounting evidence suggests that social groups are increasingly polarized in voting behavior, but how reliable are *demographic* labels as predictors of vote choice? We test for demographic sorting, using public opinion surveys between 1952–2020 and applying tree-based machine learning models to calculate out-of-sample predictions of presidential voting decisions. We calculate predictions based on voters' demographics and then gradually incorporate more information to test whether the electorate is becoming more predictable. Demographics alone typically can predict 63.5% of vote choices correctly. But contrary to the sorting hypothesis's implications, demographics have not grown more predictive over time, while partisanship has. Additional information about voters, such as issue positions or candidate perceptions, continue to be necessary for obtaining out-of-sample error rates of 5% or less. However, their added value decreases as partisanship's predictive power grows.

Keywords: Vote Choice, Elections, Polarization, Sorting, Ideology, Machine Learning, Tree-based models, Random forest

*Kim is an Assistant Professor at American University, and is the corresponding author (email: sskim@american.edu). Zilinsky is a Ph.D. candidate at New York University. We thank Jeff Gill, Shiro Kuriwaki, R. Michael Alvarez for comments.

1 Introduction

Political campaigns segment the electorate into categories based on how voters' observable characteristics are likely to correlate with voting behavior (Fenno, 1978; Hersh, 2015). To explain social and political trends, scholars also pay close attention to group behavior, and there is a deep interest in quantifying and explaining cleavages between groups. This interest is understandable: whether a readily-perceivable group, typically in terms of demographics, is a reliable base or a swing voting bloc for a particular political party has substantial implications for representation. But how reliable are demographic labels in predicting a presidential vote choice? Moreover, in light of growing polarization, do demographic groups increasingly vote in predictable ways?

After the 2016 presidential election, the media and academia both extensively discussed whether the existing political cleavages between demographic groups were widening. In the aftermath of Donald Trump's unexpected victory in 2016, researchers have sought and proposed explanations of the winner's appeal, often zooming in on the voting blocs that supported Trump, such as white voters without a college degree (Abramowitz and McCoy, 2019). Many of the post-election explanations focused on the deepening partisan divide by demographics such as race or education (Porter, 2016; Lamont et al., 2017; McQuarrie, 2017; Morgan and Lee, 2018).

Several factors are believed to have produced loyal voting blocs with few remaining persuadable voters; that partisan polarization, sorting, and the modern American partisan education gap by party have created an electorate that behaves in predictable ways. This has substantial implications for representation. Per the typical rational choice model, campaigns will focus on catering to persuadable voters rather than pandering to an electorate that will, conditional on turning out, have predictable ballot choices. If a voting bloc is "too reliable," a normative concern is that the group's interests will not be represented adequately relative to its size and importance.[1] Predictability is directly linked to the leverage that a voting bloc has over parties and candidates.

Two main ideas have been referred to as 'sorting.' *Ideological sorting*, which is the more common definition, is that operational ideology (i.e., issue positions) and partisanship are increasingly correlated (Levendusky, 2009; Hetherington, 2009; Weber and Klar, 2019).[2] *Social sorting* has been defined as a convergence of social identities and par-

[1]For example, the Democratic party has been criticized for not giving priority to issues that matter to Black voters, a key voting bloc that delivered the Biden victory (Scott, 2020).

[2]Webster and Abramowitz (2017) also report that party-line polarization is heightening for social welfare issues.

tisan identities, such as race, sexual orientation, religion, partisan factions, occupations, social movements, and so on (Mason, 2016; Mason and Wronski, 2018; Mason, 2018a). Both phenomena are believed to entrench partisanship, thereby decreasing the scope for voter persuasion.

There is little doubt that the ideological distance between Democrats and Republicans has grown (Webster and Abramowitz, 2017). Accordingly, intra-party heterogeneity has decreased; for example, Norris and Inglehart (2019) observe that "the two major parties gradually shifted to become more homogeneous internally in their cultural positions and more polarized between parties." Consequences include a greater partisan animosity (Iyengar and Westwood, 2015; Abramowitz and Webster, 2016; Bougher, 2017; Christenson and Weisberg, 2019; Iyengar et al., 2019), an increase in straight-ticket voting in recent decades (Jacobson, 2017; Burden and Kimball, 2009),[3] and an unprecedented partisan gap in presidential approval rates (Jacobson, 2019).

On the other hand, is social sorting also increasing? We narrow the definition of social sorting to *demographic groups* such as age and race, and investigate *demographic sorting*. Is the link between demographic characteristics and political opinions/partisanship tightening so that belonging to a particular group predicts voting decisions with higher accuracy? For instance, Sides (2017) write that "[t]he Democratic Party has an increasing advantage among nonwhite people. Among Hispanics, Democrats outnumbered Republicans by 23 points in 2002 but 36 points in 2016." The reduction of the white population and increasing racial diversity have been used to project the advent of the Democratic party (Teixeira et al., 2015). Abramowitz and McCoy (2019) conclude that Trump's 2016 campaign slogan of "Make America Great Again" has successfully pulled the white working-class, especially those without a college degree, away from the Democratic party.

Although hinting at it, few studies have shown whether the demographics' predictive power has consistently increased over time. In this paper, we first test whether demographic markers are increasingly more informative of vote choice, based on whether voting decisions can be accurately inferred. For this, we use random forests on data from ANES (1952–2016), CCES (2008–2018), and Nationscape surveys (July 2019–June 2020). Contrary to the prevailing narrative, we show that demographics have not become more prognostic of vote choice over time.

Second, we ask whether other data about voters increase the predictability of voting de-

[3]This is despite the gradual rollback of straight-ticket voting options in election administration (NCSL, 2020).

cisions, as implied by the findings on swing voters' disappearance (Panagopoulos, 2020). With the same data and methods, we systematically investigate how well vote choice can be inferred from various combinations of voters' characteristics. We confirm that the (1) predictive power of partisanship is increasing, and (2) while other variables additionally contribute to increasing the prediction accuracy, the added value when partisanship is already accounted for is decreasing. These results, in line with the numerous results showing increasing partisan polarization, lend credibility to our initial results.

Finally, we ask whether demographics remain informative predictors when other variables are accounted for, and if so, which ones. We use permutation-based variable importance calculations to determine whether randomly changing the values of a variable of interest significantly reduces the predictive accuracy, thus determining how 'important' a variable is. We find that being a Black voter is a top 10 variable even when party ID and issue positions are accounted for. Other variables such as age and education appear less consistently among the top predictors. However, when all variables are used for prediction, demographics completely disappear from top 10 variables, giving way to other variables such as beliefs, perceptions, and issue positions.

2 Literature and Hypotheses

Demographic attributes can naturally translate into social identities and, according to the Columbia School of political behavior (Lazarsfeld et al., 1944), they are reinforced by social networks. Membership in social groups is then said to explain politically relevant beliefs and behaviors, and voting can be viewed as akin to cultural activities which "have their origin in ethnic, sectional, class, and family traditions" (Berelson et al., 1954). In this framework, conformity is the norm.

Moreover, political elites have driven stronger group-party alignment in recent decades, making party positions clearer and more distinct to voters (Levendusky, 2009). In response, cross-cutting ties between groups have been decreasing (Mason, 2018b), increasing social polarization. Voters are expected and pressured to vote in line with their perceived group interests and against members of a disliked out-group.

It then does seem natural to state that group membership should be increasingly informative of vote choice. Essentially, this is the wide-spread assumption that "the link between demographic traits and political orientation is so strong that increases in the share of voters from demographic groups associated with support for the Democrats produce proportionate increases in Democratic support," (a view summarized but not endorsed by Shaw and Petrocik, 2020). Is this assumption true? Drawing on the existing literature, we will derive three hypotheses for the sorting thesis' testable implications.

2.1 Group-based Voting

Although social identities that align with political ideology encompass many categories such as religion and interest groups (Abramowitz and Saunders, 2008; Levendusky, 2009), our focus is on the following five demographic markers: race, education, income, age, and gender. Individuals from opposing parties now differ more on average in political opinions and their observable demographic characteristics.

Race. Although the Democratic Party has had a stable advantage among Black voters for several decades now, there are reasons to expect that the signal from a voter's race is larger than in the past. First, Trump's victory in 2016 was a continuation of "the decades-long expansion of Republican support among white working-class Americans" (Carnes and Lupu, 2020). In addition, the GOP is believed to have activated white identity (Tesler, 2016; Sides et al., 2017, 2019) in response to Barack Obama's electoral wins. In case of the Latino voters, the Democratic support varies by ethnicity and generation (Abrajano and Alvarez, 2012), while the Asian Americans are, overall, not well courted by either party (Wong et al., 2011). But racial minorities on average may have a weaker incentive to vote Republican. Note also that when the Tea Party emerged in 2010, 80–90 percent of its supporters were white (Williamson et al., 2011), strengthening the hypothesis that prediction accuracy based on race could be increasing.

Education. Three decades' worth of public opinion demonstrates that those with higher educational attainment increasingly associate with the Democratic party (Pew Research Center, 3 20). Conversely, Republicans have been gaining support among those citizens who do not have a college degree in the last decade.[4] The partisan education gap reached its peak in 2016, but note that the relationship between education and voting is sensitive to the inclusion of other variables in a model (Schaffner et al., 2018).

Income. Income at the individual level predicts vote choice, but there is some disagreement whether class-based voting has been stable (Gelman et al., 2010) or increasing over

[4]Throughout the 1990s, the Republican Party did not yet have a lead among white registered voters who were high school graduates. This group of voters was still evenly split between the two major parties.

time (Stonecash, 2000; Bartels, 2006; McCarty et al., 2008).[5] Preferences and voting are typically aligned with people's economic self-interest—for example, Ansolabehere et al. (2006) document that "the difference in the rate of Republican voting between an economic Conservative and an economic Liberal is 31 percentage points." However, the extent of the importance of economic issues for voting continues to be debated.[6] In the 2016 presidential election, the income effect is believed to have interacted with education. Carnes and Lupu (2020), for example, show that the diploma divide in 2016 "was driven largely by more affluent Americans."[7]

Age. Young people tend to lean liberal and support Democratic or progressive candidates. In the 2016 presidential popular vote, the vote margin of Democratic minus Republican votes was 24 for Millennials and 28 for Generation Z (Griffin et al., 2020). Higher age, conversely, is correlated with conservatism and voting Republican.[8] Consider also that when respondents were allowed to select up to two groups with which they have most common interests and concerns in a November 2020 YouGov poll, the most frequently mentioned category was "people in the same age group as you," followed by "people in the same political party."[9]

Gender. Voting patterns in Exit Polls suggest that men are more likely to vote Republican, but in models that control for sexist attitudes, gender does not appear to predict vote choice (Bracic et al., 2019). At the same time, gender interacts with race. For example, Junn (2017) reports that white women voted Republican in 2016 in line with their behavior in prior elections.

[5]Perhaps the best-known argument comes from McCarty et al. (2008, p. 75) who argue that there has been growing "stratification of partisanship by income," with high-income voters increasingly voting Republican.

[6]The relationship between income and Republican partisanship at the individual level, while robust nationally, is moderated by local context (especially ethnic composition) according to detailed analyses of voter files (Hersh and Nall, 2016).

[7]Tree-based methods are ideally suited for identifying interactive relationships between variables and exploiting them to produce accurate predictions.

[8]Williamson et al. (2011) found that at least 75% of Tea Party supporters were over 45 years old.

[9]That is, rather than inferring the importance of group memberships, respondents were asked directly: "Would you say that you share a lot of common interests and concerns with other people of people who are [SAME GROUP], or would you say that age is not really relevant?". In this context, respondents suggested that class ("people who have about the same amount of money as you"), ethnicity, and geographic proximity were less indicative of common interests than age (YouGov, 2020).

2.2 Hypotheses

Our objective is to test whether demographic sorting is taking place. We argue that the testable implication of demographic sorting is increasing predictability of demographics on presidential vote choice. Based on the summarized relationships between demographics and political behavior, we derive the following hypothesis:

> *Hypothesis 1 (Increasing Demographic Sorting): Vote choice will become increasingly predictable based on voters' demographic features alone (with other information about voters withheld).*

To give credibility to our results, we also show the well-established result of ideological polarization. Noting that an explicit self-reported party label should be a stronger signal, we propose the next hypothesis:

> *Hypothesis 2 (Increasing Party ID Sorting): Including explicit party ID will make predicting voting decisions increasingly easy over time, and accuracy will be higher relative to sparser models such as using only demographics.*

Finally, inasmuch parties are seen as ideological brands (Woon and Pope, 2008) which own certain issues (Egan, 2013) we propose the final hypothesis:

> *Hypothesis 3 (Sufficiency of Party ID): Beyond the initial sets of features (party ID and demographics), other voter characteristics, such as issue positions, will contain minimal diagnostic information about vote choice.*

2.3 Machine Learning and Political Behavior

There are several reasons for using supervised machine learning methods to test the above hypotheses, especially the tree-based methods that we choose. First of all, the metrics we use to evaluate the results are performance-based on correct out-of-sample predictions. Second, random forests allow flexible interaction structures between variables, uncovering hidden relationships in large datasets (Montgomery and Olivella, 2018). Third, when the set of potential predictors is large, researchers can prune the set of covariates or identify the most important predictors in distinguishing the outcome variable, instead of arbitrarily restricting the set of allowable model specifications (Kim et al., 2020).

We emphasize the first upside of regression trees—and machine learning in general—relative to the family of parametric models usually employed in social science, typically under the maximum likelihood umbrella. A common approach in the existing literature is to estimate a set of logistic regressions and evaluate their performance based on the percent of correctly classified observations (or McKelvey-Zavoina's pseudo R^2) *in-sample*. However, when the out-of-sample fit is not reported, readers cannot evaluate whether the reported models overfit to the given sample.

Several recent papers, recognizing these advantages, have used these flexible non-parametric methods to explore complex structures in political behavior and perform direct prediction. For example, Bonica (2018) and Bonica and Li (2021) use them to predict legislators' behaviors and issue positions based on campaign contribution records, while Kim et al. (2020) identifies the best predictors of turnout. As the aforementioned authors, we rely on tree-based methods to achieve the best possible prediction given the covariates.

Finally, we wish to derive variable importance measures by permutation (Breiman, 2001) to identify which variables contribute the most to increasing accuracy, instead of relying on statistical significance. The variable importance measures how much the prediction accuracy decreases when a given variable is either removed or reshuffled. This metric is not available under the traditional regression framework where the judgment about variable importance generally involves comparing coefficient magnitudes or significance-based criteria.[10]

Note that our paper is attempting to harness the power of nonparametric modeling of pure prediction algorithms for the best classification under the training/testing paradigm. The intent is to assess the accuracy of both long-standing and potentially short-term predictions in a given election cycle and to see if trends are consistent over time. We apply random forests across all data and across model specifications to take advantage of these characteristics consistently, insofar as the number of covariates p in the richest models approaches the number of observations n in one of our datasets.[11] For a comprehensive 'checklist' of differences between regression models and prediction algorithms, see Table 5 in Efron (2020).

[10]It is important to note that statistical significance does imply that a variable is diagnostic or predictive (Lo et al., 2016).

[11]We compare the performance of these models to logistic regression and classification trees in the Appendix.

3 Data and Methodology

We use three sets of public opinion surveys: American National Election Studies (1952–2016, every four years), Cooperative Congressional Election Study (2008–2018, every two years), and UCLA Nationscape surveys (50 weekly waves in 2019 and 2020). The target variable for prediction is presidential vote choice, which is self-reported,[12] subsetted to respondents who voted for either a Democratic or a Republican candidate.[13] Using all these major surveys in public opinion will help us validate the results and check that our results are not the product of a single survey.

To predict voting decisions, we use random forests, a method for aggregating predictions from regression and classification trees. An individual tree is estimated by sequentially splitting the data on the basis of an optimally chosen cut-off point of the most informative variable.[14] To remove excessive dependence of tree structures on the algorithmic decisions early in the splitting process, a subset of observations and predictors is drawn each time a new classification tree is estimated. An aggregation of tress corrected for inter-tree correlation is the random forest (RF).

To investigate the extent to which voting behavior is inferrable based on voters' observable characteristics, we use four nested variable specifications for the analysis. Table 1 shows the labels of and the variables included in each specification. Naturally, the third and the fourth specifications will usually consist of imperfectly overlapping sets of variables for each survey/wave. We include these specifications for benchmark purposes, given that the survey questionnaires reflect the issue cleavages of the day, such as the Iraq war or the Affordable Care Act.

	Specification Label	Variables Included
1	Demographics Only	Gender, race, age, income, education
2	Demo. + PID	Specification 1 + 7-point Party ID
3	Demo. + PID + Issues	Specification 2 + All issue-related questions
4	All Covariates	Specification 3 + All other questions

Table 1: Four Nested Specifications and Corresponding Variables

[12] If there is a post-wave and a pre-wave, we use the post-wave variable. For the CCES mid-term election waves, we use the previous election cycle's presidential vote choice. For Nationscape, we use the respondents' vote intention for the 2020 presidential election.

[13] For ANES, the cumulative dataset was used. For the CCES dataset, seven waves from 2008 and 2018 were used, after extensive wrangling and coding of equivalent variables. The cumulative CCES content was not available at the time of the analysis.

[14] A variable is chosen in a given step if using that variable minimizes deviance.

All categorical variables are converted into dummy variables, including a variable to represent nonresponse missing values. Variables with near-zero variance at the 1% level or variables with more than twenty different responses, such as ZIP codes, are dropped to guard against too much sparsity. Only clearly continuous variables—such as age, number of children, or amount donated to political campaigns—are kept as continuous.[15] Nonresponses are treated as a separate category instead of listwise deletion if the variable is categorical.

After cleaning, the data is split into training and testing datasets with an 80:20 ratio. Using the `caret` and the `ranger` package in `R`, we run a class prediction via random forests (Breiman, 2001; Kuhn, 2008; Wright and Ziegler, 2017). All code is publicly available at a GitHub repository: https://github.com/sysilviakim/surveyML. For comparison purposes, we also run a logistic regression and a CART model, the results of which are available in the Appendix.

Note that while we have placed results from different surveys side by side for comparison, we do not claim that they are by default comparable. These surveys were designed each for their respective purposes, and the number of respondents and survey modes differ. Even within the ANES survey, the survey modes have undergone some changes (e.g., the addition of the web mode). In particular, questions on cleavage issues are usually different.

Unlike ANES or CCES, Nationscape is a high frequency (online) poll where the number of respondents in a typical week is 6,250. We use the data collected prior to the onset of the COVID pandemic because of concerns about changes to the sample composition during an economic crisis. We randomly draw 20% of the pooled dataset, yielding a sample of 25,937 for the training set. We thus deliberately maintain a sample size between that of ANES and the CCES. We then draw 5,187 respondents from the hold-out set and evaluate the models' performance on this set of respondents.

As we have stated in Section 2, we argue our approach has several compelling aspects: better performance, a rich set of interactions between variables that can be flexibly explored, derivation of variable importance measures, and guarding against overfitting

[15] We chose to encode ordered categorical variables (such as education and party ID) as sets of binary features—so that, for example, a seven-point party ID becomes six binary variables—for several reasons. First, this allows us to be consistent over various surveys and years. Second, we avoid assuming that items measured on an ordered scale have linear or additive effects. Finally, if ordered categorical variables are treated as continuous, item nonresponses would be dropped but one-hot encoding allows us to keep all survey respondents even when their responses contain missing values, because a separate binary variable is created (e.g. for respondents with a missing income category).

by evaluating the performance of models out-of-sample. Again, the last aspect yields honest estimates of model performance.[16] Second, regardless of model complexity, by focusing on solving a prediction problem, we can transparently summarize each model with metrics such as out-of-sample accuracy and the area under the receiver operating characteristic (ROC) curve (AUC).

There are also concerns and criticisms associated with borrowing methods from computer science, including low interpretability and higher computation time and costs. The former concern can be partially mitigated by storing intermediate outputs or visualizing how much predictions can be attributed to individual model inputs. The latter issue is becoming less severe as the availability of high-performance computing (HPC) improves, but we recognize that a lack of access to HPC can be a barrier introducing resource-based inequities across researchers.

4 Results

4.1 How Much Can Demographics Alone Predict Vote Choice?

Before we investigate the evidence for our hypotheses, we first quantify how much, on average, demographics can predict presidential vote choice. Figure 1 shows the time-series plots of out-of-sample accuracy values over time for all three surveys. The top panel shows accuracy rates over time for models estimated using only information on respondents' gender, race, education, income, and age.

We find that when using just demographics, the accuracy for vote choice predictions is generally low, typically less than 65%. More specifically, the average across all surveys and waves is 63.5%. When determined for individual surveys, it is 63.1% for ANES, 64.7% for CCES, and 63.4% for the Nationscape's 2019–20 waves. It is striking to see that these numbers are remarkably similar even across different survey administrations, varying sampling methods, and in different time periods.

Is 63.5% a high number? It is certainly better than random guesses of the two-party vote choice, which on average would be 50%. But is it high enough, considering the emphasis that the literature places on demographic variables? Compared to the spotlight they receive, altogether the accuracy does not seem extraordinarily high. The next question is then whether, per the first hypothesis, the predictiveness of demographics has increased

[16] Examples of earlier work employing similar methods in political science include Samii et al. (2016), Kim et al. (2020), and Demir et al. (2021).

over time.

4.2 No Evidence of Demographic Sorting Over Time

Hypothesis 1 (Increasing Demographic Sorting). Can demographic labels alone predict presidential vote choice better in the polarized era compared to the past? The time trends shown in the first panel of Figure 1 do not suggest that this is the case. Test-set accuracy from 1952 to 2020 is remarkably stable.

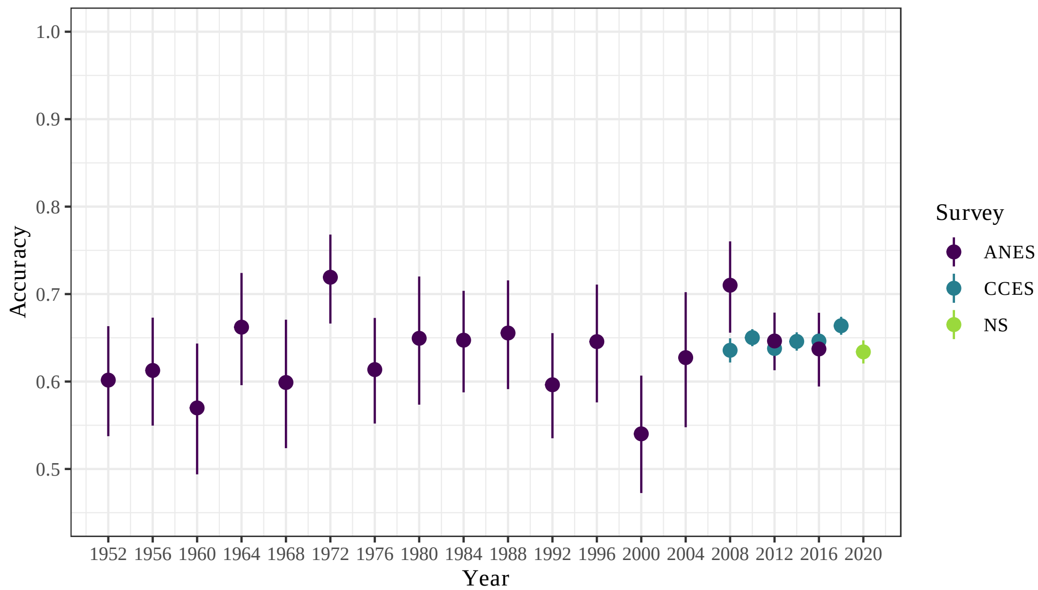
Two notable exceptions are the elections of 1972 and 2008. For both the Nixon vs. McGovern case and the Obama vs. McCain contest, the ANES-based accuracy is slightly above 70% (71.9% in 1972 and 71.0% in 2008). However, CCES-based accuracy in 2008[17] was only 63.6% and, crucially, not significantly different than the accuracy rates we observe in 2016 (64.6%) or 2020 (63.4% in the Nationscape data). Again, overall, given the area covered by the 95 percent confidence intervals of accuracy, we see that predictions are typically only 10 to 15 percentage points better than random guesses. In fact, in 1960 and 2000, predictions were only marginally better than predictions obtained by chance (respectively 57% and 54% accuracy).

A simple regression slope of accuracy on years with just ANES data is 0.0004 with a standard error of 0.0006 (p-value of 0.46). The slope, while positive, is not statistically significant, thereby providing no evidence to reject the null. If all surveys are pooled—again, with the caveat that accuracy between different surveys may not be directly comparable—the slope is 0.0004 with a standard error of 0.0004 (p-value of 0.24).

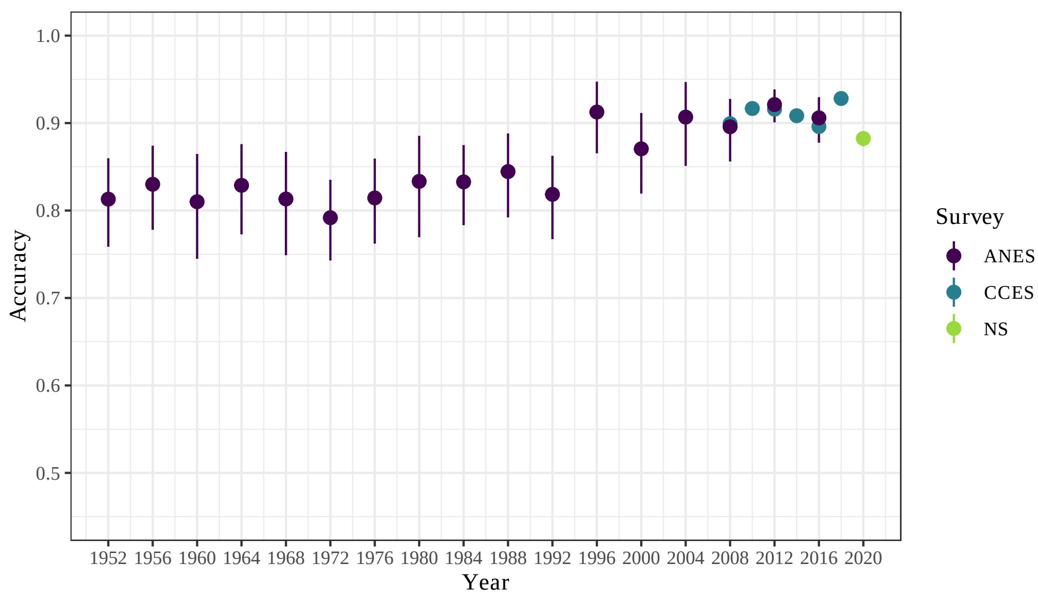
Thus, there is no evidence to suggest that there is demographic sorting—voters' demographic characteristics do not provide more informative signals for vote choice over time. The results suggest that the electorate has not become more polarized along demographic lines in a way that is *informative about voting behavior*.[18] Note that this is not about a specific demographic group (e.g., white working-class men in their 30s) but the overall prediction ability of combinations of demographics, which may explain the disparity between the results of this paper and the literature's recent focus.

[17] Note that the large confidence interval of ANES survey datasets are due to their relatively small size, compared to CCES or Nationscape.

[18] Note that AUC, over time, seems to be increasing unlike accuracy, with ANES-only slope of 0.0022 (standard error 0.0005, p-value of 0.0008), (pooled slope of 0.002 (standard error 0.0004, p-value of < 0.001) hinting that over all possible threshold values, the ability to separate the Democratic vote from the GOP vote has increased over time, starting in the 1970s. But when classifying vote choices, researchers rarely use a threshold other than 50%.



(a) Demographics Only



(b) Demo. + PID

Figure 1: Accuracy 95% Confidence Interval, Presidential Vote Prediction Over Time, Demographics Only and Demographics and Party ID. ANES (1952–2016), CCES (2008–2018), and Nationscape (2020). Predictions are evaluated on the hold-out sample.

4.3 Prognostic Power of Party ID and Partisan Polarization

Hypothesis 2 (Increasing Party ID Sorting). Next, we turn to tests of our second hypothesis that stated that partisan identification would become more informative over time, consistent with the existing literature. Models that take advantage of data on respondents' explicit partisan identification, measured on a 7-point scale, are summarized in the bottom panel of Figure 1. We also present Figure 2 which shows point estimates of three performance metrics (accuracy, AUC, and the F-1 score), which also displays these performance metrics for the remaining nested specifications. The full set of performance metrics is available in the Appendix.

We find that partisanship, jointly with basic demographics, has indeed become a significantly more prognostic variable over time. Before the 1992 election, PID-based accuracy (together with accuracy) never exceeded 85%. However, starting with Bill Clinton's re-election, the same specification generally classifies at least 90% of voting decisions correctly. The linear regression slope testing for a temporal trend in the rates displayed in the bottom panel of Figure 1 is 0.0018 with a standard error of 0.0003 (p-value < 0.001), which is in terms of effect size more than four times the coefficient size from the first specification reported above.[19]

In addition to accuracy, Figure 2 also displays the evolution of the AUC (a metric that guards against class imbalance) and the F-score (which balances precision and recall) over time. We see that—for party-inclusive specifications—both of these metrics are increasing, confirming the patterns mentioned above. We thus find empirical support for our second hypothesis, which aligns with the well-established results of an increasingly polarized U.S. electorate.

To be sure, the predictions based on the second specification leave on the table other knowable attributes of voters. We explore the implications of their inclusion, which brings us to our final hypothesis.

Hypothesis 3 (Sufficiency of Party ID). While partisanship is now more prognostic of vote choice, other factors continue to be important in the sense of providing additional useful information for inferring respondents' vote choice, as we can see in Figure 2. More specifically, on average, accuracy of ANES predictions improves by 0.7 percentage points when issue variables are added on top of demographics and party ID. In addition,

[19] Pooled results yield a slope of 0.0018 with a standard error of 0.0002 (p-value < 0.001).

accuracy improves by 8.2 percentage points once all other variables have been added.[20] Some examples of variables included in the fourth and final specification are non-policy opinions. For example, the top variable in terms of permutation importance for the 2018 CCES was a belief that Trump colluded with Russia to influence the 2016 election, and for ANES 2016, it was the perception about whether honesty well describes the Democratic presidential candidate.

Hence the evidence generally does not favor the third hypothesis. The patterns uncovered by these models suggest that it is possible to glean significant information about voters' behavior even after accounting for their partisanship. Views on policy issues consistently reveal more information about behavior, above and beyond partisanship. Moreover, other questions asked on public opinion surveys (occupation, subjective class identification, group attitudes, political knowledge, media consumption, beliefs, perceptions, and so on) still contain a significant amount of additional information that can be used to improve predictions about voters' behavior.

However, it is clear that in recent years, the value-added from the set of all variables in terms of prediction is, on average, decreasing. Once party ID is accounted for along with demographics, the ability of other variables to be put to good use in better predicting vote choice is more limited compared to the era with lower mass-level polarization.

4.4 Variable Importance of Demographics

Given Section 4.2, we explore a related question: do demographics remain as top important variables when other variables are accounted for? If so, which ones? We calculate this with permutation-based variable importance measures. Table 2 shows which demographic characteristics remain as top 10 variables when either (1) only party IDs are included (Specification 2), or (2) on top of that, issue variables are also included (Specification 3). The variables that appear in the same row are aligned by importance from left to right.

The demographics quickly give way to party ID in variable importance and further disappear once issues are accounted for. The only variable that consistently stays as informative about vote choice is identifying as Black. Age and education are somewhat important but are less significant and less consistent over different surveys and waves.

[20]When pool accuracy rates from all surveys, they improve by 1.3 percentage points when issue variables are added on top of demographics and party ID, and by 7.2 percentage points when all other variables have been added.

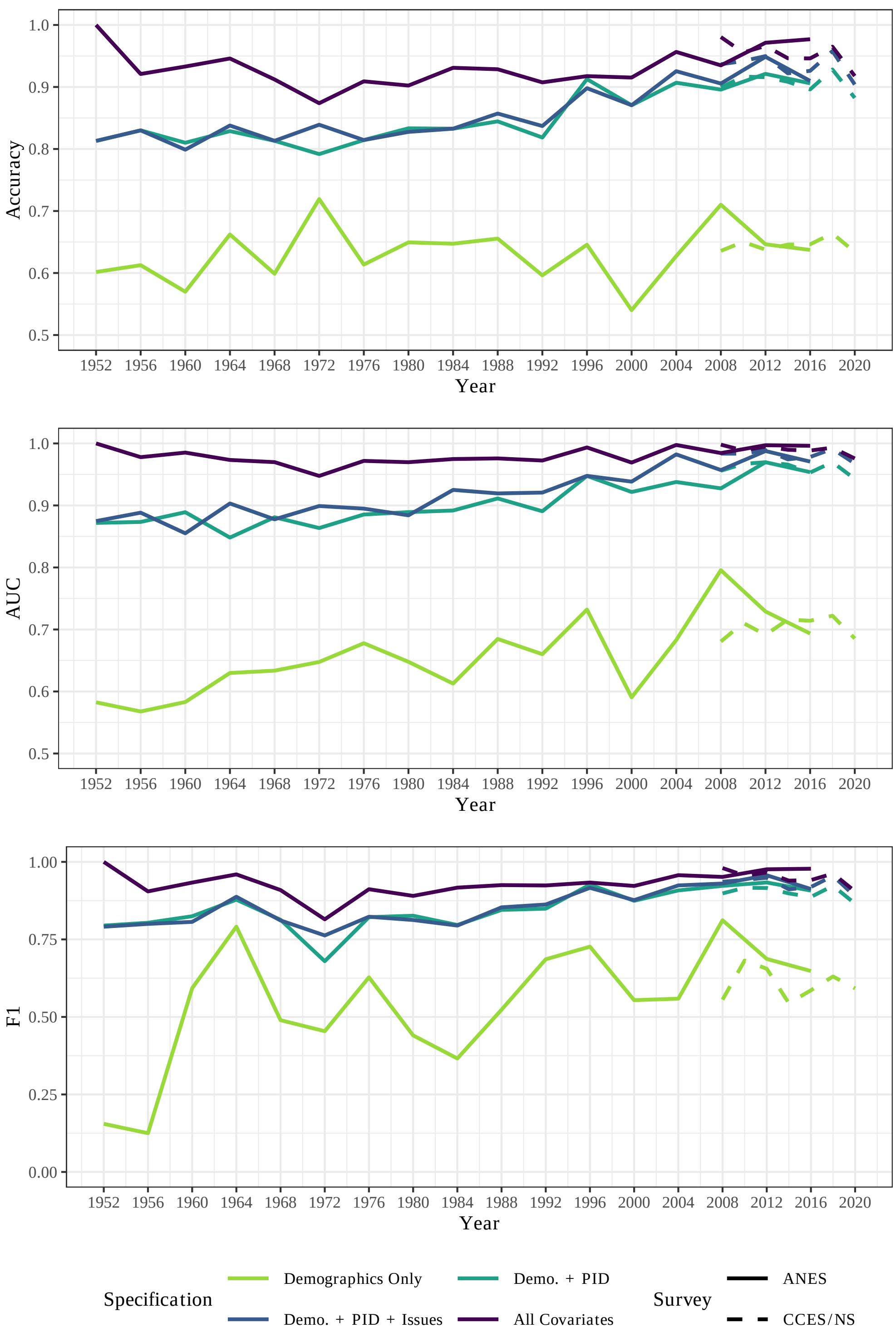


Figure 2: Performance of Presidential Vote Prediction Over Time, All Surveys, Random Forests, Accuracy/AUC/F1 Scores

Year	V1	V2	V3	V4	V5
1952	Black	Some college	2-year college		
1956	Income: 68-95 %tile	Income: refused	High school graduate	Age	
1960	Income: 96-100 %tile	Age	Income: 68-95 %tile	Income: 34-67 %tile	
1964	Black	2-year college	Age	Income: 34-67 %tile	Income: 68-95 %tile
1968	Black	Some college	High school graduate	Age	Income: 34-67 %tile
1972	Black	Age	Income: 34-67 %tile	2-year college	
1976	Black	Age	Income: 68-95 %tile	2-year college	
1980	Black	Income: 96-100 %tile	Hispanic	Age	2-year college
1984	Black	High school graduate	2-year college	Age	
1988	Black	Hispanic	Income: 68-95 %tile	Gender	
1992	Black	2-year college	High school graduate	Age	
1996	Black	2-year college	Hispanic	Gender	
2000	Black	Income: 96-100 %tile	2-year college	High school graduate	
2004	Black	Age	2-year college	Some college	
2008	Black	Hispanic	2-year college	High school graduate	
2012	Black	Hispanic	2-year college	Age	
2016	Black	2-year college	Some college	Hispanic	

(a) PID Included (ANES)

Year	V1	V2
1952	Black	
1956	Income: 68-95 %tile	
1960	Age	
1964		
1968	Black	Age
1972	Black	
1976	Black	
1980	Black	
1984	Black	
1988	Black	
1992	Black	
1996		
2000		
2004		
2008	Black	
2012	Black	
2016	Black	

(b) PID/Issues Included (ANES)

Year	V1	V2	V3	V4
2018	Black	Age	Post-grad	High school graduate
2016	Black	Post-grad	High school graduate	4-year college
2014	Black	Age	High school graduate	4-year college
2012	Black	Age		
2010	Black	High school graduate	4-year college	Some college
2008	Black	Age	High school graduate	4-year college

(c) PID Included (CCES)

Year	V1	V2
2018	Black	
2016	Black	
2014	Black	Age
2012	Black	
2010	Black	
2008		

(d) PID/Issues Included (CCES)

Table 2: Demographics Remaining in Top 10 Variables By Variable Importance, Presidential Vote Choice, Random Forests, ANES (1952–2016) and CCES (2008–2018)

Identifying as Black consistently retains strong prediction power on vote choice even after accounting for party IDs.

However, note that in Specification 4 (all covariates), none of the variables remain in the top 10 variables. With ANES data, you can expand the threshold up to the top 15 variables, whereby for 1972 and 2008, identifying as Black will still be counted as one of the more important variables. With CCES data, the threshold needs to be expanded to the top 30 variables, whereby for 2014 and 2018 being Black emerges in the top important variables set. Given the results in Section 4.2, we could say that identifying as Black is a strong predictor in the sparse covariate set, but not *growing* stronger over the years.

5 Conclusion

Demographic attributes can function as markers of social groups, and membership in these social groups does, to an extent, carry political meaning. Voters' demographic characteristics help improve vote choice predictions compared to random guesses—given the five demographic characteristics, the probability of an accurate prediction with a random forest model is on average 63.5%. This, although higher than 50%, suggests that for most people, memberships in their income group, age group, gender, education group, or even ethnic group is not politically 'sorted' strongly enough to translate to particularly accurate signals about their voting decisions.

Moreover, the accuracy of predicted vote choice inferred based on voters' demographic attributes has not grown over the years. Our findings based on three sets of national surveys from 1952 to 2020 show that out-of-sample prediction accuracy has not significantly increased over the last seven decades. Our results are therefore not consistent with the first hypothesis: demographic sorting is not increasing.

We validate our first result by showing results from four nested model specifications, increasingly incorporating more information about respondents. In line with existing findings showing higher polarization, we show that once partisanship is no longer withheld from the set of predictors, we do observe, as expected, a massive increase in accuracy. Furthermore, inferring vote choice with just the combination of demographics and party ID grows easier over time.

We also note that predictions based on partisanship and demographics can further be improved by including issue positions from each survey. Adding extra variables (including non-policy features such as voters' political knowledge, media consumption, attitudes tapping into identity considerations, and other survey instruments) to the set of model features also generally yields higher accuracy. While not fully comparable due to questionnaires changing over time, the full models generally perform well, lending credibility to the method of choice. However, once partisanship has been accounted for, the added value from additional variables is decreasing over time. In recent survey waves, there is only minimal gain from richer specifications besides partisanship and demographics.

To check that demographics are not vital to the prediction, we also check permutation-based variable importance measures. Only the strongest demographic signal of identifying as Black consistently remains among the top 10 important variables in specifications that include partisanship and/or issues. However, in the full model, where all possible covariates are used for prediction, no demographic features appear as a top 10 variable.

Considering all this, we conclude that demographic sorting has not translated into voting behavior. Without information about respondents' partisanship, even sophisticated random forest models typically only achieve out-of-sample accuracy of up to 65%, and this performance metric has not increased over the last seventy years. As the predictive power of party ID grows stronger, it dominates the signal from other covariates, diminishing their marginal predictive power. Therefore, while our results validate scholarly findings on ideological sorting and polarization, we find no support for vote-based demographic sorting.

References

Abrajano, Marisa and R. Michael Alvarez (2012). *New Faces, New Voices: The Hispanic Electorate in America*. Princeton University Press.

Abramowitz, Alan and Jennifer McCoy (2019). United States: Racial Resentment, Negative Partisanship, and Polarization in Trump's America. *The ANNALS of the American Academy of Political and Social Science 681*(1), 137–156.

Abramowitz, Alan I. and Kyle L Saunders (2008, April). Is polarization a myth? *The Journal of Politics 70*(2), 542–555.

Abramowitz, Alan I. and Steven Webster (2016, March). The Rise of Negative Partisanship and the Nationalization of U.S. Elections in the 21st Century. *Electoral Studies 41*, 12–22.

Ansolabehere, Stephen, Jonathan Rodden, and James M Snyder (2006, June). Purple America. *Journal of Economic Perspectives 20*(2), 97–118.

Bartels, Larry M (2006). What's the Matter with What's the Matter with Kansas? *Quarterly Journal of Political Science 1*(2), 201–226.

Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee (1954). *Voting: A Study of Opinion Formation in a Presidential Campaign*. University of Chicago Press.

Bonica, Adam (2018). Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning. *American Journal of Political Science 62*(4), 830–848.

Bonica, Adam and Zhao Li (2021). Inferring Candidates' Issue-Specific Positions from Itemized Campaign Contributions Using Supervised Machine Learning.

Bougher, Lori D. (2017, September). The Correlates of Discord: Identity, Issue Alignment, and Political Hostility in Polarized America. *Political Behavior 39*(3), 731–762.

Bracic, Ana, Mackenzie Israel-Trummel, and Allyson F Shortle (2019). Is sexism for white people? gender stereotypes, race, and the 2016 presidential election. *Political Behavior 41*(2), 281–307.

Breiman, Leo (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Burden, Barry C. and David C. Kimball (2009). *Why Americans Split Their Tickets: Campaigns, Competition, and Divided Government*. University of Michigan Press.

Carnes, Nicholas and Noam Lupu (2020). The White Working Class and the 2016 Election. *Perspectives on Politics*.

Christenson, Dino P. and Herbert F. Weisberg (2019, October). Bad Characters or Just More Polarization? The Rise of Extremely Negative Feelings for Presidential Candidates. *Electoral Studies 61*, 102032.

Demir, Mehmet Özer, Biagio Simonetti, Murat Alper Başaran, and Sezgin Irmak (2021, January). Voter Classification Based on Susceptibility to Persuasive Strategies: A Machine Learning Approach. *Social Indicators Research*, 1–16.

Efron, Bradley (2020). Prediction, Estimation, and Attribution. *Journal of the American Statistical Association 115*(530), 636–655.

Egan, Patrick J. (2013). *Partisan priorities: How issue ownership drives and distorts American politics*. Cambridge: Cambridge University Press.

Fenno, Richard F. (1978). *Home Style: House Members in Their Districts*. Longman.

Gelman, A, L Kenworthy, and Yu-Sung Su (2010, December). Income inequality and partisan voting in the United States. *Social Science Quarterly 91*(5), 1203–1219.

Griffin, Rob, William H. Frey, and Ruy Teixeira (2020). America's electoral future: The coming generational transformation. Center for American Progress.

Hersh, Eitan D. (2015, June). *Hacking the Electorate: How Campaigns Perceive Voters*. Cambridge University Press. Google-Books-ID: DEuqCQAAQBAJ.

Hersh, Eitan D and Clayton Nall (2016, April). The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records. *American Journal of Political Science 60*(2), 289–303.

Hetherington, Marc J (2009, April). Putting Polarization in Perspective. *British Journal of Political Science 39*(2), 413–448.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science 22*(1), 129–146.

Iyengar, Shanto and Sean J. Westwood (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science 59*(3), 690–707.

Jacobson, Gary C. (2017, April). The Triumph of Polarized Partisanship in 2016: Donald Trump's Improbable Victory. *Political Science Quarterly 132*(1), 9–41.

Jacobson, Gary C. (2019, March). Extreme Referendum: Donald Trump and the 2018 Midterm Elections. *Political Science Quarterly 134*(1), 9–38.

Junn, Jane (2017). The trump majority: white womanhood and the making of female voters in the u.s. *Politics, Groups, and Identities 5*(2), 343–352.

Kim, Seo-young Silvia, R. Michael Alvarez, and Christina M Ramirez (2020, February). Who voted in 2016? using fuzzy forests to understand voter turnout. *Social Science Quarterly 45*(1), 5–11.

Kuhn, Max (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software 28*(1), 1–26.

Lamont, M, Bo Yun Park, and Elena Ayala-Hurtado (2017). Trump's electoral speeches and his appeal to the American white working class. *The British Journal of Sociology 68*(S1), S154–S180.

Lazarsfeld, Paul F., Bernard Berelson, and Hazel Gaudet (1944). *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Duell, Sloan and Pearce.

Levendusky, Matthew (2009). *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. Chicago: University of Chicago Press.

Lo, A, H Chernoff, T Zheng, and Shaw-Hwa Lo (2016, December). Framework for making better predictions by directly estimating variables' predictivity. *Proceedings of the National Academy of Sciences 113*(50), 14277–14282.

Mason, Lilliana (2016). A Cross-Cutting Calm: How Social Sorting Drives Affective Polarization. *Public Opinion Quarterly 80*(S1), 351–377.

Mason, Lilliana (2018a). Losing Common Ground: Social Sorting and Polarization. *The Forum 16*(1), 47–66.

Mason, Lilliana (2018b). *Uncivil Agreement: How Politics Became Our Identity*. University of Chicago Press.

Mason, Lilliana and Julie Wronski (2018). One Tribe to Bind Them All: How Our Social Group Attachments Strengthen Partisanship. *Political Psychology 39*(S1), 257–277.

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal (2008). *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.

McQuarrie, Michael (2017). The revolt of the Rust Belt: place and politics in the age of anger. *The British Journal of Sociology 68*(S1), S120–S152.

Montgomery, Jacob M and Santiago Olivella (2018). Tree-based models for political science data. *American Journal of Political Science 62*(3), 729–744.

Morgan, Stephen and Jiwon Lee (2018). Trump Voters and the White Working Class. *Sociological Science 5*, 234–245.

NCSL (2020). Straight Ticket Voting States. https://www.ncsl.org/research/elections-and-campaigns/straight-ticket-voting.aspx.

Norris, Pippa and Ronald Inglehart (2019). *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. Cambridge University Press.

Panagopoulos, Costas (2020). *Bases Loaded: How US Presidential Campaigns Are Changing and Why It Matters*. Oxford University Press.

Pew Research Center (2018-03-20). Wide Gender Gap, Growing Educational Divide in Voters' Party Identification.

Porter, Eduardo (2016). Where were trump's votes? where the jobs weren't. New York Times. December 13.

Samii, Cyrus, Laura Paler, and Sarah Zukerman Daly (2016, October). Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia. *Political Analysis 24*(4), 434–456.

Schaffner, Brian F, Matthew MacWilliams, and Tatishe Nteta (2018, March). Understanding white polarization in the 2016 vote for president: The sobering role of racism and sexism. *Political Science Quarterly 133*(1), 9–34.

Scott, Eugene (2020). Black voters delivered Democrats the presidency. Now they are caught in the middle of its internal battle. *Washington Post*.

Shaw, Daron and John Petrocik (2020). *The Turnout Myth*. Oxford University Press.

Sides, John (2017). Race, religion, and immigration in 2016: How the debate over american identity shaped the election and what it means for a trump presidency. *Democracy Fund Voter Study Group*.

Sides, John, Michael Tesler, and Lynn Vavreck (2017). How Trump Lost and Won. *Journal of Democracy 28*(2), 34–44.

Sides, John, Michael Tesler, and Lynn Vavreck (2019, August). *Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America*. Princeton University Press. Google-Books-ID: pgqSDwAAQBAJ.

Stonecash, Jeff (2000). *Class And Party In American Politics*. Routledge.

Teixeira, Ruy, William H. Frey, and Rob Griffin (2015). States of Change: The Demographic Evolution of the American Electorate, 1974–2060. Technical report, Center for American Progress.

Tesler, Michael (2016, April). *Post-Racial or Most-Racial?: Race and Politics in the Obama Era*. University of Chicago Press. Google-Books-ID: GWmkCwAAQBAJ.

Weber, Christopher and Samara Klar (2019). Exploring the Psychological Foundations of Ideological and Social Sorting. *Political Psychology 40*(S1), 215–243.

Webster, Steven W and Alan I. Abramowitz (2017, June). The ideological foundations of affective polarization in the u.s. electorate. *American Politics Research 45*(4), 621–647.

Williamson, Vanessa, Theda Skocpol, and John Coggin (2011, March). The Tea Party and the Remaking of Republican Conservatism. *Perspectives on Politics 9*(1), 25–43.

Wong, Janelle S., S. Karthick Ramakrishnan, Taeku Lee, Jane Junn, and Janelle Wong (2011). *Asian American Political Participation: Emerging Constituents and Their Political Identities*. Russell Sage Foundation.

Woon, Jonathan and Jeremy C. Pope (2008). Made in congress? testing the electoral implications of party ideological brand names. *Journal of Politics 70*(3), 823–836.

Wright, Marvin N. and Andreas Ziegler (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software 77*(1), 1–17.

YouGov (2020). HuffPost: Common Interests.