Predicting the 2018 United States House of Representatives Elections and Voting Patterns with Polling Data

Group: Debbie Chung, Wilson Lui, Carolyn Morris, Kejia Shi[1]
Mentors: David Rothschild[2], Tobias Konitzer[3], Tian Zheng[4]

May 7, 2018

Abstract

Nationwide surveys are widely adopted for gathering household information, evaluating public policies, and predicting elections; however, no unofficial survey can achieve census-level coverage. With survey data gathered by PredictWise using Pollfish, a mobile survey platform, we build a multilevel regression and poststratification (MRP) model to predict the two-party vote shares for each of the 435 congressional districts and the District of Columbia in the forthcoming 2018 United States House of Representatives Elections. We then use these predictions to investigate the changes in voter turnout in specific demographics that would be needed to change the balance of power in the House of Representatives. In addition to widely-used demographic and geographic information, we incorporate responses to psychometric survey questions using three weighting schemes to evaluate their effects on the model. We identify several question topics that improve our model's predictive accuracy and find evidence that adding multiple topics simultaneously produces approximately linear improvements in accuracy.[5]

1. Introduction

The 2018 U.S. House of Representatives Elections will be closely watched as a test of how the political landscape has changed since the election of Donald Trump in 2016. In particular, survey data used to predict the outcome of the election will be closely scrutinized. Surveys remain one of the primary tools for forecasting elections, yet levels of public mistrust in surveys are higher than ever due to some high-profile misses during the 2016 Presidential Election. These misses can be partly traced back to non-representative survey sampling that failed to accurately capture voter turnout (the population of people who actually voted). In this report, we make use of multilevel regression and poststratification (MRP), which can mitigate some of the issues with non-representative survey data by adjusting the survey population to be more representative of the voter turnout population. This method is especially effective at dealing with demographic groups that are traditionally underrepresented in polling data.

There are a number of issues encountered in traditional survey methodology that have made forecasting elections difficult in recent years. One difficulty involves the shifts from landline to cell phone to smartphone. The dominant survey channel over the past couple of decades has been random digit dialing (RDD), in which landline telephone numbers are randomly dialed. As households move away from landlines, RDD is no longer a reliable method of collecting randomized responses. Fortunately,

---

[1] The authors would like to be considered to have contributed equally. Data Science Institute, Columbia University in the City of New York, {ddc2129, wl2522, cm3491, ks3403}@columbia.edu;
[2] Microsoft Research (New York), davidmr@microsoft.com;
[3] Department of Communication, Stanford University, tobias.konitzer@gmail.com;
[4] Department of Statistics, Columbia University in the City of New York, tian.zheng@columbia.edu;
[5] All code related to this project can be found at https://github.com/wl2522/MSR2

smartphones have enabled new methods of reaching respondents. Our project utilizes a mobile survey platform called Pollfish to gather survey data. A second difficulty is nonresponse. Response rates have fallen from 36% in 1997 to 9% in 2016[6]. The response population is different from the nonresponse population, creating a nonresponse bias. It has been demonstrated that nonresponse can have large effects on survey results (Lohr, 2009). There are many established ways to deal with this potential nonresponse bias. How do we obtain an unbiased estimate despite the presence of nonresponsive people?

Simply ignoring the nonresponse leaves the unmeasurable bias intact. In order to reduce such bias, weighting adjustments can be used to make respondents and nonrespondents in the same class similar. Similar to weighting adjustment, poststratification is another way to compensate for nonresponse (Little, 1993). We modify the weights based on population counts so that the sample is calibrated to the true population counts in the poststrata. Raking, first used in the 1940 Census, is an advanced poststratification method used when poststrata are formed using more than one variable, but only the marginal population totals are known (Deming and Stephan, 1940). Specifically, raking repeatedly estimates weights across each variable in turn until the weights converge, forcing the survey totals to match the known population totals. It requires an additional assumption: the response probabilities depend only on the row and column and not on the particular cell (Lohr, 2009). Nevertheless, raking has its difficulties in practice. First, the algorithm may not converge if some of the cell estimates are zero. This will happen inevitably if we want to include many variables of interest. The other notable downside of raking is the "overadjustment" danger. If there is little relation between the extra dimension in raking and the cell means, raking can increase the variance rather than decrease it. In the voting preference estimation, it is likely that cell sparsity and the inclusion of new dimensions can create serious issues, as we need to match selected demographic and geographical information to perform accurate adjustments.

The state-of-the-art methodology used in national/subnational survey data analysis is multilevel regression and poststratification (MRP). First developed by Gelman and Little (1997), MRP is widely adopted for static evaluation of polling data to recover state-level estimates of public opinion. It is especially useful when there is a small number of observations in certain states. Park et al. (2006) take geography into account to estimate state-level opinions from national surveys. MRP can be used to estimate opinions in other subnational areas, such as congressional districts (Kastellec et al., 2016).

MRP is a two-step process: first, build a Bayesian hierarchical model and then poststratify the results. A Bayesian hierarchical model improves over frequentist statistics in many ways (Gelman et al., 2014). In short, Bayesian modeling treats the parameters as random variables and uses subjective information to establish assumptions on these parameters. A hierarchical model captures the structure of observations. This is suitable when observations fall into a number of clusters and the distribution over outcomes is determined jointly by a shared parameter across clusters and individual parameters shared among observations within a cluster, which may be different across clusters. A hierarchical model has proven to be robust, with the posterior distribution being less sensitive to the more flexible hierarchical priors. After building a Bayesian hierarchical model, poststratification corrects for clustering and other statistical issues that may bias estimates obtained from nonresponse.

MRP has been shown to produce reasonably accurate estimate of state public opinion using as little as a single large national poll with only around 1,400 respondents. (Lax and Phillips, 2009) Also, respondents from Alaska and Hawaii can be included in national polls. Unlike raking or simple weight adjustment, MRP deals with sparsity introduced when clustering respondents by demographic and geographic information. When assuming random effects, MRP improves a model's explanatory power and produces insights on how much of public opinion is determined by these features.

---

[6] http://www.pewresearch.org/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/

In this report, we use MRP to build the major party voter turnout and Democratic Party preference models in order to predict the two-party vote shares for each of the 435 congressional districts and the District of Columbia in the forthcoming 2018 United States House of Representatives Elections. With these models as a baseline, we achieve 75.5% and 86.5% out-of-sample predictive accuracy, respectively. Section 2 describes the dataset and our findings from performing exploratory data analysis. Section 3 summarizes our baseline results. Section 4 proposes possible extensions to the baseline models. Specifically, we include psychometric variables based on responses to survey questions as additional features. We convert these responses using different weighting schemes and find that the best one produces an average increase of 0.75% in accuracy in the Democratic Party preference model. Section 5 considers future work, the limitations of our model, and possible ethical concerns.

2. Dataset

The complete polling dataset contains two parts: 1) answers to questions concerning political opinions and psychometric variables, and 2) the respondent's demographic and geographic information. The survey data was collected during five months: October and November 2017 and January, February and March 2018. For each month, there are two psychometric topics investigated, including populism, racial resentment, traditionalism, compassion, globalism, economic populism, authoritarianism, trust in institutions, and climate change. All months have the same political opinion questions and we are mainly concerned with their voting preference in the 2018 Midterm Elections, specifically "Who will you vote for in the House of Representatives in 2018?". The total number of valid responses after preprocessing is 20,922. The monthly counts of responses increase every month, while in October 2017 we only have 2,878 responses and in March 2018 we have 7,432 responses.

The poststratification space is a prebuilt dataset that estimates the space of likely voters for the 2018 US House of Representatives Election. The space is separated into different cells based on various demographic variables.[7]

One of the key issues that MRP tackles is the imbalance of the dataset in terms of demographic and geographic information, which may create significant and unmeasurable bias if we use traditional weighting strategies. The imbalances in our dataset is demonstrated in Figure 1. About 60% of the respondents are female, compared to 52% of the electorate. Compared to the two electorate structures, the compositions of age and marital status in the polling dataset are totally different. The polling dataset has more young (25 - 34) and unmarried people compared to typical election spaces that skew older (over 54) and married. In the state-level distribution plot, we see that people from Michigan and Texas are overrepresented in the polling data; Texas is predicted to have much fewer people voting in 2018.

For our prediction model, we encode the response to the question "Who will you vote for in the House of Representatives in 2018?" as our dependent variable. This variable is visualized in Figure 2. In terms of the proportion of responses on the y-axis, Republicans are favored by voters who are white or over the age of 45. By comparing the blue lines, representing Democratic votes, and the red lines, representing Republican votes, we further find that, in general, the respondents who are male, older, white, unmarried, and non-college educated constitute larger proportions of the likely Republican voters.

---

[7] The 2018 likely voter projection space was provided by the mentors at the congressional district level across demographics (age, gender, race, education, marital status), geography (state, district, urbanicity), party affiliation and psychometrics (authoritarianism). The total count of likely voters is predicted to be 122,665,900. A variant model using the Bayesian Item Response Theory (Bayesian IRT) is built to include the psychometric dimension. Further details on this procedure will be provided in their upcoming white paper.

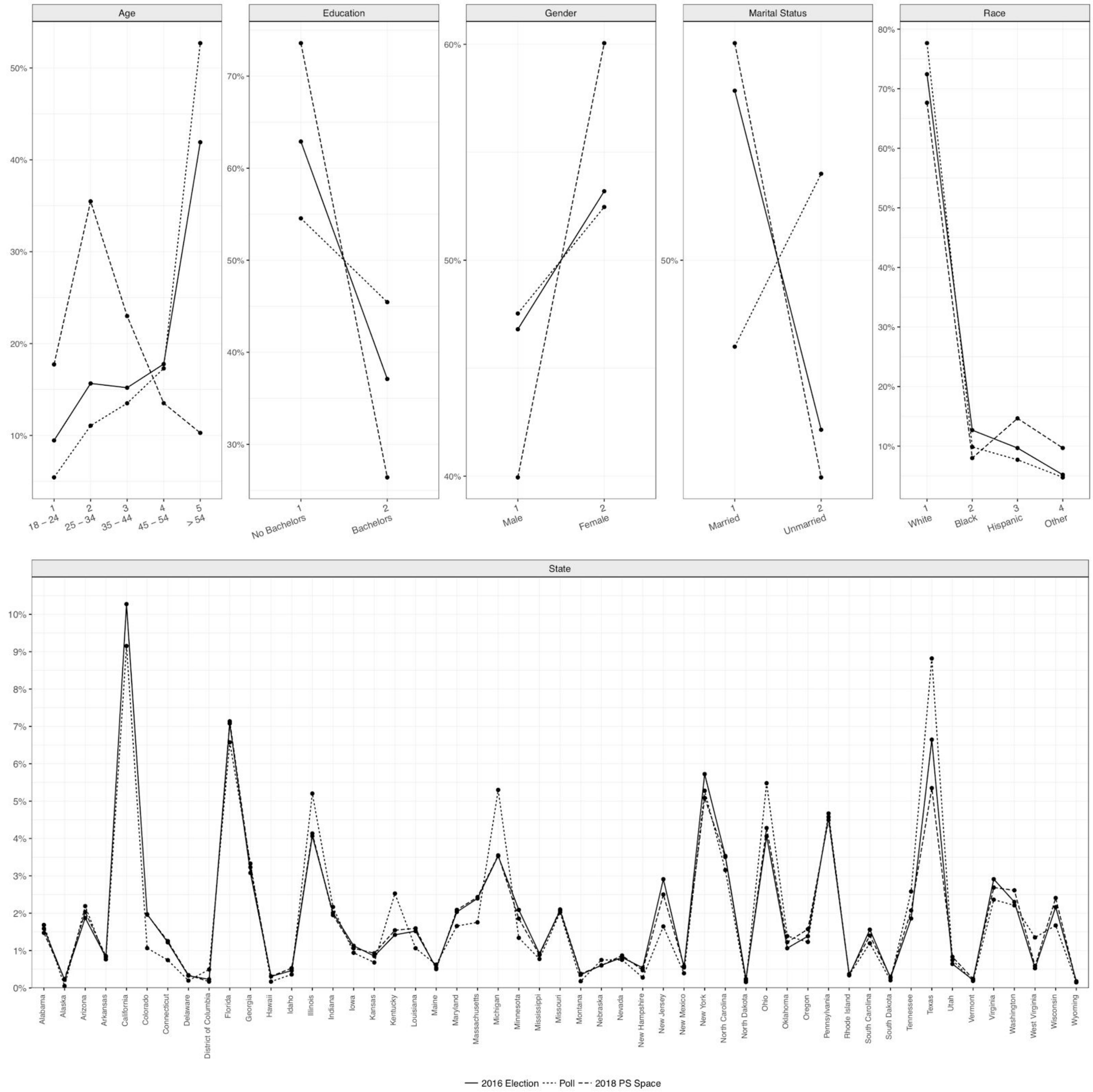

Figure 1. Distributions of demographics from the Pollfish polling data and the predicted 2018 likely voter poststratification space compared to voter turnout in the 2016 Presidential Election

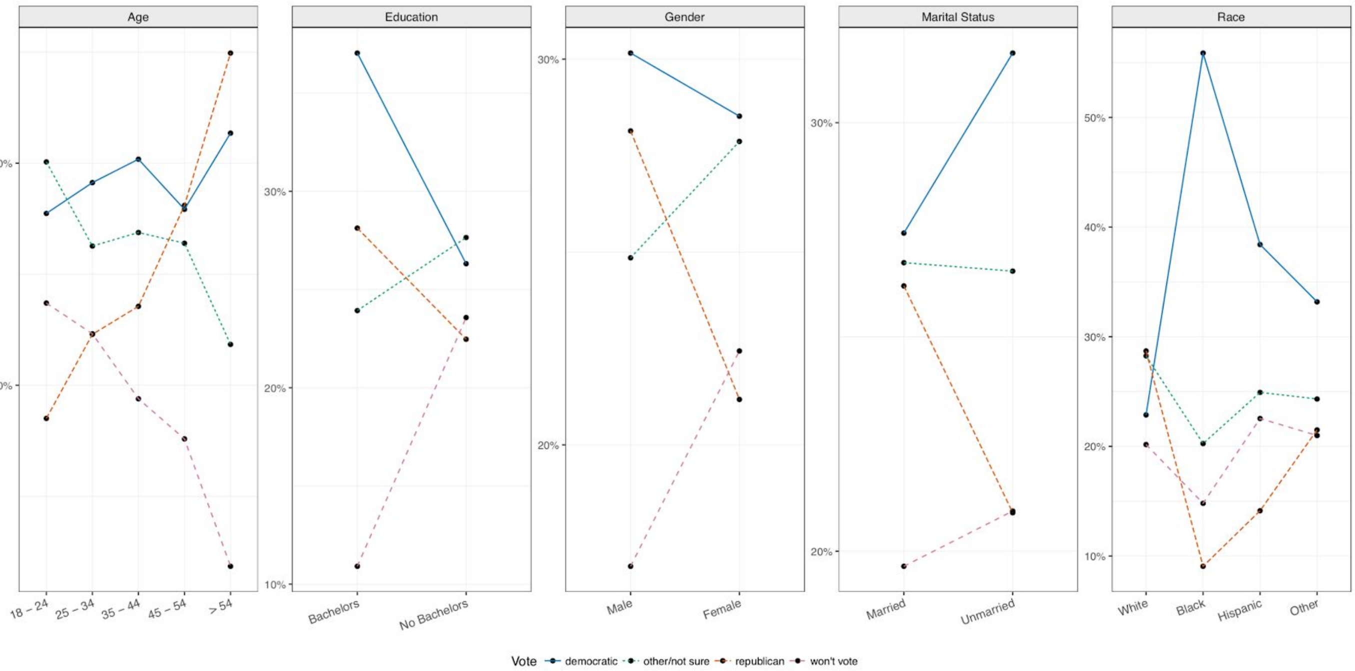

Figure 2. Distributions of responses to the question "Who will you vote for in the House of Representatives in 2018?" among selected demographics (column values sum up to 1).

3. Two-Party Vote Share Prediction Model

3.1 Data Preprocessing

We coded demographic information into categorical variables and match this coding to the poststratification space. Additional data cleaning steps are performed. Specifically, levels of race, education, and marital status are combined and responses from people under the age of 18 are removed. Additional preprocessing steps are taken to account for missing geographical information (state, congressional district, and zip code).[8] The survey data has no missing values in terms of demographic (age, gender, race, education, party affiliation) information. There are instances of the same respondent taking the survey at different times. For the purpose of training our model, we remove repeat responses from the same respondent, keeping only the most recent one.

3.2 Random Effects

Our baseline prediction model is a Bayesian hierarchical logistic regression model that attempts to learn the voting behavior of people belonging to certain subpopulations defined by different combinations of five demographic variables: 1) age, 2) gender, 3) race, 4) education level, and 5) party affiliation. These five variables were selected based on historical polling research, which showed that these five variables and the two-way interactions between them are correlated with voting behavior. In

[8] We combine state-level responses into one district for the following states, which only have one congressional district: Alaska, Delaware, Montana, North Dakota, South Dakota, Vermont, and Wyoming. For the other states, we map the zip code to the corresponding congressional district based on publicly available data. Uniformly random assignment is performed for respondents living in congressional districts that map to multiple zip codes. These zip code to district mappings were obtained from: https://github.com/OpenSourceActivismTech/us-zipcodes-congress.

addition, our model also learns the voting behavior of congressional districts as a whole, which has an effect on how people living in those districts tend to vote. Specifically, our model uses a non-centered parameterization and weakly informative priors, such that for each variable α,

$$\mu_\alpha = 0 + \Delta_\alpha \sigma_\alpha$$
$$\Delta_\alpha \sim \mathcal{N}(0,1)$$
$$\sigma_\alpha \sim Exp(1)$$

3.3 District-Level Effects

Since the survey data does not provide any information on how each district votes, we augment our model with three external sources of information for each district: 1) the Census Division[9] corresponding to the state that district belongs in, 2) the log odds of Donald Trump's two-party vote share in the 2016 Presidential Election[10], and 3) the Cook Partisan Voting Index[11] (CPVI). Thus, each of the 435 districts and the District of Columbia, represented by $\alpha_{district}$, has a mean, $\mu_{district}$, defined such that

$$\alpha_{district} \sim \mathcal{N}(\mu_{district}, 1)$$
$$\mu_{district} = \alpha_{division} + \beta_{trump} logit(vote_{2016}) + \beta_{cook} cook_{district}$$
$$\alpha_{division} \sim \mathcal{N}(0,1)$$
$$\beta_{trump} \sim \mathcal{N}(0,1)$$
$$\beta_{cook} \sim \mathcal{N}(0,1)$$

The Census Divisions are included due to the historical observation that neighboring groups of states tend to vote together in the same direction. The district-level 2016 Presidential Election results directly provide information on each district's most recent voting behavior. The CPVI measures how much more each district leans towards a major party compared to the national average. This metric is derived from the district-level results of the past two presidential elections. The CPVI is not provided for the District of Columbia due to the fact that their congressional representative is a non-voting delegate. We therefore calculate its CPVI using the 2012 and 2016 Presidential Election results[12]. Combining the district-level effects and the random effects, our model is defined as

$$\eta = \beta_0 + \alpha_{age} + \alpha_{gender} + \alpha_{race} + \alpha_{education} + \alpha_{party} + \alpha_{age\times gender} + \alpha_{age\times education} + \alpha_{age\times party} + \alpha_{age\times race} + \alpha_{gender\times education} + \alpha_{gender\times party} + \alpha_{gender\times race} + \alpha_{education\times party} + \alpha_{education\times race} + \alpha_{party\times race} + \alpha_{district}$$

[9] https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
[10] https://www.dailykos.com/stories/2016/11/25/1601042/-Nerd-Alert-This-spreadsheet-contains-every-presidential-election-by-state-from-1828-to-2016
[11] http://cookpolitical.com/index.php/introducing-2017-cook-political-report-partisan-voter-index
[12] https://docs.google.com/spreadsheets/d/1MNl4B-tQW6uIMmVmabmbL1bUMsDscp9AvXLREFI7swk/edit#gid=415911445

Our baseline model is actually a combination of two models: a major party vote turnout model, which predicts whether a survey respondent will vote for a major party in the 2018 Midterm Elections, and a Democratic Party preference model, which predicts whether a respondent will vote for the Democratic Party in the 2018 Midterm Elections given that he or she will be voting for a major party. The output of each model corresponds to the probability of a person belonging to a particular subpopulation and living in a certain district voting for the Democratic Party in the 2018 Midterm Elections. We then perform poststratification in order to project those probabilities onto the space of likely voters. This is done using the following formula, introduced in Park et al. (2006):

$$p_d = \frac{\sum_{j \in d} N_j \pi_j^{major} \pi_j^{dem}}{\sum_{j \in d} N_j \pi_j^{major}}$$

Each model's output, π_j, represents a vector where each value represents a probability associated with a specific subpopulation j. N_j represents a vector containing the population counts for each j in district d. In order to obtain the predicted two-party vote share for the Democratic Party for each d, we weigh the products of the two probabilities obtained for each j by the number of people belonging to j who live in d and taking the sum across all j in d.

3.3 Model Evaluation Methodology

We evaluate the predictive accuracy of each model using cross-validation, training the models on 80% of the data and testing on the remaining 20%. We remove respondents from the test set if they are also present in the training set. When training the Democratic Party preference model, we restrict the dataset to only contain respondents who stated that they will vote for a major party.

Model	Test Set Prediction Accuracy
Major Party Voter Turnout	0.755
Democratic Party Preference	0.865

Table 1: The out-of-sample predictive accuracy achieved by each model using cross-validation.

3.4 Model Predictions and Prediction Space Visualization

Our model predicts that in the upcoming 2018 Midterm Elections, the Democrats will win 174 out of the 436 congressional districts, resulting in a net loss of 19 seats that they currently hold. This contradicts the current predictions of the vast majority of pollsters[13]. We discuss potential reasons for our model's Republican bias in section 5.2.

[13] As of May 7th, 2018: https://projects.fivethirtyeight.com/congress-generic-ballot-polls/?ex_cid=rrpromo

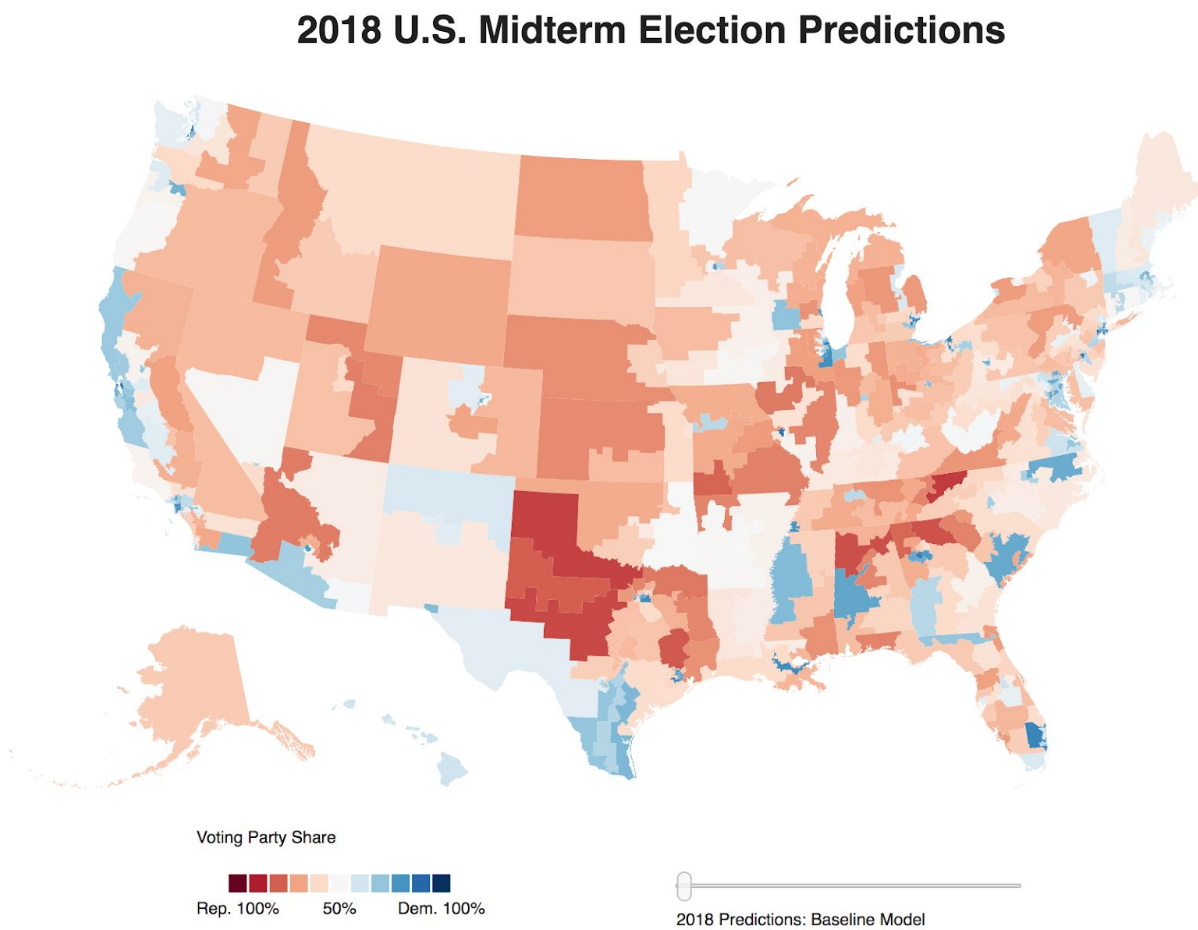

Figure 3: A visualization of our model's predicted outcomes for the 2018 Midterm Election[14].

4. Extensions

4.1 Psychometric Variables

We extend the prediction model described in section 3.1 by incorporating psychometric variables that were indirectly measured by questions included in the survey. The overall dataset used in section 3.1 is comprised of five subsets of approximately 6,000 survey responses each. Each subset of respondents was asked a mutually exclusive set of questions. On average, each topic contains seven questions. For each topic, we add the variable α_{topic} to our model:

$$\eta = \beta_0 + \alpha_{age} + \alpha_{gender} + \alpha_{race} + \alpha_{education} + \alpha_{party} + \alpha_{age \times gender} + \alpha_{age \times education} + \alpha_{age \times party} + \alpha_{age \times race} + \alpha_{gender \times education} + \alpha_{gender \times party} + \alpha_{gender \times race} + \alpha_{education \times party} + \alpha_{education \times race} + \alpha_{party \times race} + \alpha_{district} + \alpha_{topic}$$

[14] https://bl.ocks.org/carolynamorris/raw/18e6fbf579cb9c539c5c5b15a7b17de5/

The variable α_{topic} represents a respondent's belief or opinion on a given topic or the presence or absence of a certain personality trait (we will henceforth refer to such beliefs, opinions, and personality traits as "traits" and the presence of a trait as "identifying with" a trait). The questions were included in the surveys with the intent that these psychometric variables would provide additional information on voting behavior and preferences. The questions were designed to ask about the respondent's opinions in an indirect manner in order to elicit a more honest response, due to the fact that directly asking whether or not the respondent identifies with a certain trait may cause offense or embarrassment, such as when asking a respondent whether he or she has feelings of racial resentment or a lack of compassion. For many questions, this is done by asking whether the respondent agrees or disagrees with a particular statement.

Survey #	Topic	Example Question
1	Populism	"The system is stacked against people like me."
2	Racial Resentment	"Racial minorities can overcome prejudice without any special favors."
4	Traditionalism	"More influence of churches in daily life would make society better."
4	Compassion	"Homelessness is sometimes the right price to pay for lacking work ethic."
5	Economic Populism	"Too many millionaires and billionaires lead to inequality hurting people like me."
5	Globalism	How do you feel about rolling back free trade agreements?
6	Authoritarianism	Which quality is more important for children: independence vs. respect for elders?
6	Trust in Institutions	Do you trust what the FBI and CIA reports?
7	Presidential	Compared to other recent presidents, is President Trump a moral leader?
7	Climate	How do you feel about government spending and regulations to address climate change?

Table 2: The list of topics (and the survey in which they appeared) and examples of the corresponding survey questions. Note that we omit survey questions pertaining to globalism from this analysis due to concerns about how those questions were worded.

4.1.1 Model Evaluation Methodology

We incorporated each of these psychometric variables into our model and compare the predictive accuracy with that of our existing baseline model. Once again, we remove repeat responses from the same respondent, keeping only the most recent one. Note that for each topic, we only use the subset of survey data which includes questions pertaining to that topic, resulting in a much smaller dataset. Moreover, we retrain our baseline model with the same subset of data in order to obtain a fair comparison. We also experimented with including pairs of psychometric variables that appeared together in the same survey to explore how this affects the accuracy of the model when compared to including each variable separately.

In order to convert a respondent's answers into observed values for α_{topic}, we must decide how to weigh the answer choices for each question. We tested three different weighting schemes[15]. For each combination of topic and weighting scheme, we train the major party turnout and Democratic Party preference models on 80% of the survey data and evaluate them using the remaining 20%.

[15] Refer to the appendix for an explanation of each weighting scheme.

4.1.2 Psychometric Variable Model Results

In general, the major party turnout model was largely unaffected by the psychometric variables, with the exception of “Trust in Institutions”, which produced modest improvements of approximately 1%. This variable measures how trusting of institutions a respondent is, regardless of where on the political spectrum those institutions fall (for example, Fox News vs. the New York Times). Thus, how trusting of institutions a person is may be indicative of whether or not that person will vote for a major party.

Topic	Major Party Turnout Model				Democratic Party Preference Model			
	Baseline	Scheme 1	Scheme 2	Scheme 3	Baseline	Scheme 1	Scheme 2	Scheme 3
Populism	0.740	0.740 (0.000)	0.739 (-0.002)	0.740 (0.000)	0.874	0.872 (-0.002)	0.872 (-0.002)	0.871 (-0.004)
Racial Resentment	0.740	0.740 (0.000)	0.740 (0.000)	0.737 (-0.004)	0.874	0.895 (0.021)	0.899 (0.024)	0.886 (0.011)
Populism + Racial Resentment	0.740	0.741 (0.001)	0.741 (0.001)	0.739 (-0.002)	0.874	0.893 (0.019)	0.899 (0.024)	0.889 (0.015)
Traditionalism	0.752	0.753 (0.001)	0.758 (0.006)	0.754 (0.002)	0.865	0.882 (0.017)	0.876 (0.011)	0.882 (0.017)
Compassion	0.752	0.751 (-0.001)	0.758 (0.006)	0.750 (-0.002)	0.865	0.865 (0.000)	0.874 (0.009)	0.871 (0.006)
Compassion + Traditionalism	0.752	0.756 (0.004)	0.758 (0.006)	0.758 (0.006)	0.865	0.878 (0.013)	0.882 (0.017)	0.872 (0.007)
Economic Populism	0.787	0.788 (0.001)	0.790 (0.004)	0.788 (0.001)	0.869	0.883 (0.014)	0.891 (0.022)	0.888 (0.019)
Authoritarianism	0.728	0.728 (0.000)	0.728 (0.000)	0.728 (0.001)	0.853	0.866 (0.013)	0.864 (0.011)	0.872 (0.019)
Trust in Institutions	0.728	0.731 (0.004)	0.740 (0.012)	0.738 (0.010)	0.853	0.851 (-0.002)	0.842 (-0.011)	0.851 (-0.002)
Authoritarianism + Trust	0.728	0.728 (0.001)	0.748 (0.020)	0.737 (0.009)	0.853	0.864 (0.011)	0.870 (0.017)	0.875 (0.022)
Presidential	0.781	0.786 (0.005)	0.780 (-0.002)	0.777 (-0.005)	0.898	0.886 (-0.011)	0.881 (-0.017)	0.881 (-0.017)
Climate	0.781	0.781 (0.000)	0.780 (-0.002)	0.783 (0.002)	0.898	0.900 (0.003)	0.900 (0.003)	0.895 (-0.003)
Presidential + Climate	0.781	0.781 (0.000)	0.777 (-0.005)	0.773 (-0.008)	0.898	0.884 (-0.014)	0.886 (-0.011)	0.889 (-0.008)

Table 3: The out-of-sample accuracies obtained by each combination of model, topic, and weighting scheme. The difference in accuracy between each model and its baseline are included in parentheses. Cells highlighted in green indicate models that obtained at least 1% higher accuracy than the baseline, while those in red obtained at least 1% lower accuracy than the baseline.

Several psychometric variables appear to have an effect on the predictive accuracy of the Democratic Party preference model. Specifically, the addition of "Racial Resentment", "Traditionalism", "Economic Populism", and "Authoritarianism" yielded slight improvements in our model's ability to predict whether someone will vote for the Democrats or not. These findings are aligned with academic studies that have found that a preference towards authoritarianism (Taub, 2016) and traditionalism (Chokshi, 2018) are indicative of support for Donald Trump and the Republican Party. Meanwhile, the inclusion of the "Presidential" topic, which measures a respondent's opinion of President Donald Trump, negatively affects the accuracy of the Democratic Party preference model. This may be due to the fact that this information overlaps with the CPVI and Donald Trump's two-party vote share in the 2016 Presidential Election, which are already included in our model as district-level effects.

We also found that including two psychometric variables simultaneously in the Democratic Party preference model appears to produce linear changes to the prediction accuracy, where the change is roughly equal to the sum of the changes in accuracy resulting from including each variable separately. Thus, depending on the traits in question, including two psychometric variables in the model may improve the predictive accuracy more than including only one.

Among the three weighting schemes, scheme #2 performed the best across all psychometric variables, yielding an average increase in accuracy of 0.75%. This suggests that a weighting scheme that groups answer choices in a binary manner (either an answer identifies with a trait or it does not) and describes how much a respondent identifies with a trait on an integer scale is optimal.

4.1.3 Poststratification with Psychometric Variables

Among the traits that were measured in the survey data, the poststratification space we used only includes authoritarianism. Moreover, this variable is expressed in a binary manner (either "authoritarian" or "not authoritarian"). Therefore, we can only perform poststratification using weighting scheme #1. The district-level two-party vote share predictions obtained from the authoritarianism model is extremely similar to that of the baseline model, with the root mean squared error and correlation coefficient between the two being 0.012 and 0.998, respectively. The authoritarianism model predicts that the Democratic Party will win 173 districts in the upcoming election, compared to 174 districts in our original prediction. The only district where the outcome changed is Kentucky's 6th district, where the Democratic Party vote share dropped from 51.2% to 48.8%. Therefore, it appears that like the modest gains in model accuracy, the addition of psychometric variables only produces modest changes in the model predictions.

4.2 Voter Turnout Adjustment

To analyze the impact of demographic shifts in voter turnout, we made adjustments to various cells in the poststratification space. Using a brute force method, each demographic cell (such as Black college-educated women over the age of 54) was increased by a multiplier to artificially increase the turnout in each cell while holding all other cells constant. Our goal was to determine which demographic groups might tip the balance of power towards the Democrats. By focusing on the Democrats of each race, we identified the "tipping point" at which that group would need to vote in order to increase our baseline outcome of 174 districts to 218 (the number needed for a majority). For white Democrats, the tipping point is when their turnout increases by 40%. For black Democrats, it is 266%. For Hispanic Democrats, it is 586%. For Democrats of other races, it is 890%. Our findings are visualized alongside our baseline predictions[16]. These numbers are dramatic, underscoring the distance between our predicted

[16] To see all alternate outcomes, visit https://bl.ocks.org/carolynamorris/raw/18e6fbf579cb9c539c5c5b15a7b17de5/

outcome and what is required to achieve a majority. One caveat is that these tipping points are calculated at the national level, so the increases in turnout for the Democrats would be concentrated in districts they are already likely to win. A more realistic and useful analysis can be done at the district level to find the tipping point where turnout is likely to sway competitive districts. This is an area for further study.

5. Conclusion

5.1 Summary of Results

By combining the outputs of a model that estimates voter turnout for the major parties and a model that estimates voter preference between the Democratic Party and the Republican Party and performing poststratification using an estimated space of likely voters, we predicted the two-party vote share for each congressional district in the 2018 Midterm Elections. Using cross-validation with 80% training data and 20% test data, we achieved 75.54% and 86.69% out-of-sample predictive accuracy with the major party voter turnout and Democratic Party preference models, respectively. With this model, we predicted that the Democrats will win 174 out of 436 districts in the upcoming elections.

We then extended our model to include psychometric variables, which measure how much a respondent identifies with certain personality traits or beliefs. For each psychometric variable, we compared the cross-validated accuracy of our model with and without that variable included. We found that these variables had little to no effect on the major party turnout model while producing small changes in accuracy in the Democratic Party preference model. We also experimented with adding pairs of psychometric variables to the model and found that the changes in accuracy produced by each pair were approximately equal to the sum of the changes produced by each variable separately.

Using a model that includes authoritarianism as a variable, we performed poststratification and produced updated two-party vote share predictions for each district. These predictions are quite similar to our baseline results, predicting that the Democrats will win 173 of the 436 districts.

To analyze the impact of demographic shifts in voter turnout on our model predictions, we also adjusted our poststratification space by iteratively increasing the rate of turnout by various factors. We found that when focusing only on individual demographics, very large increases in turnout at the national level would be required to change the balance of power.

5.2 Limitations and Potential Biases

Our baseline model predictions for the 2018 Midterm Elections imply that the Democrats will lose 19 seats in the House of Representatives. This deviates from the current political sentiment[17] and other polls, which predict that the Democrats will win a majority of the districts and change the balance of power. While there is the possibility that our predictions are accurate, we found that this Republican bias may be due to how the survey data was collected. Through Pollfish, smartphone users browsing on the Amazon app are presented with a pop-up ad that allows them to take the survey for a chance to win a $10 gift card. Moreover, the surveys were only conducted in English, meaning that non-English speakers, who tend to support the Democratic Party, were excluded from the survey. There is a possibility that among English speaking users of the Amazon app, those who would be enticed by the chance of winning a $10 gift card tend to be more supportive of the Republican Party. For example, this is reflected in districts of Southern California, where the Hispanic population is much larger than other districts within the same region. Specifically, no interaction term of district and race is included in our model, as the inclusion will

[17]As of May 7th, 2018

break convergence and slow down the model training significantly. We believe this may introduce a downward bias against the Democrats.

In addition, there is no way to examine the prediction accuracy of the poststratification space before the elections. As required by our methodology, poststratification adjusts the disaggregation bias with a correct space. The current poststratification space, representing our best belief of this year's voter turnout, may still look quite similar to the voter turnout data from the 2016 Presidential Election, in which the Republican Party won or maintained control of the Presidency, Senate, and House of Representatives. This may potentially create a bias towards the Republican. Further investigation of the voter turnout dynamics would be required in order to better predict the space and definitively identify the source and direction of the bias.

5.3 Ethical Issues

Ethical concerns relating to this project may arise due to the detailed and personal nature of the information collected from the surveys. In general, a large majority of the survey responses include the respondent's AdID or IDFA, which are used to identify mobile users for advertising purposes by Google and Apple, respectively. Similarly, a vast majority of the responses include the respondent's location at the zip code level. For these reasons, there may be a risk of identifiability with respect to respondents who are racial minorities and live in sparsely populated areas. For example, one respondent of Hispanic descent is located in Lewes, Delaware, a city with a current population of 2,955[18]. With only 2.84% of the residents in Lewes being Hispanic (approximately 84 people), it would not be difficult to identify this respondent's identity with public records and the demographic information that he provided in the survey, such as his gender, year of birth, and marital status. Considering the fact that some of the survey questions deal with controversial topics that a respondent may not be comfortable with discussing publicly, serious ethical issues would arise if respondents could be individually identified and connected to their responses.

We feel that these ethical concerns are not relevant to our work in this project due to the fact that our results are only reported in aggregate at the district level, rather than at the individual level. Moreover, our model does not use the AdID/IDFA data at all. In our model code, this information is removed shortly after loading the survey data. Since the survey responses and the poststratification space data are not being publicly shared, we feel that there is no risk of someone taking our district-level predictions and using them to identify individual survey respondents.

5.4 Future Work

A major limitation of the survey dataset that we used is the fact that it contains four mutually exclusive sets of questions, with none of them being asked to every respondent. This limited the ways in which we could incorporate psychometric variables in our model, such that we could add at most two of those variables to our model simultaneously. Since our results show evidence that adding several variables to the model can produce linear increases in predictive accuracy, one could further test this idea by using survey data where each respondent is asked questions concerning three or more different topics.

Similarly, we were restricted by which psychometric variables we could use to produce election predictions due to the fact that the poststratification space only included authoritarianism. It would be interesting to see if the predictions change significantly when poststratifying using a likely voter space that includes other psychometric variables that produced increases in model accuracy, such as traditionalism or racial resentment. Combined with more recent survey responses leading up to the

[18] https://datausa.io/profile/geo/lewes-de/

elections, one could investigate how the predictions change over time and whether or not the inclusion of other psychometric variables has a significant effect on the predicted outcomes.

Lastly, as our prediction is based on monthly polling data, we hope to represent the temporal dynamics of public opinion. However, building a dynamic model is not as straightforward as the static MRP models demonstrated in this report. It is still an active research area to systematically include time as a feature into the prediction. Classical time series models may fail to capture the random effect as well as MRP does, while building separate MRP models at different time points may risk losing the dynamics completely. On the other hand, while our prediction task is at the district-level, including any interaction terms between time and district may break the convergence and lower the prediction accuracy. We are actively researching novel methods to build such dynamic models.[19]

References

Lax, J. R., & Phillips, J. H. (2009). Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review*, *103*(3), 367-386.
Lohr, S. (2009). *Sampling: design and analysis*. Nelson Education.
Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, *23*(2), 127-35.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL: CRC press.
Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, *88*(423), 1001-1012.
Park, D. K., Gelman, A., & Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. *Public opinion in state politics*, 209-28.
Stephan, F. F., Deming, W. E., & Hansen, M. H. (1940). The sampling procedure of the 1940 population census. *Journal of the American Statistical Association*, *35*(212), 615-630.
Kastellec, J. P., Lax, J. R., & Phillips, J. (2010). Estimating state public opinion with multi-level regression and poststratification using R. *Unpublished manuscript, Princeton University*.
Taub, A. (2016, March 01). The rise of American authoritarianism. Retrieved from https://www.vox.com/2016/3/1/11127424/trump-authoritarianism
Chokshi, N. (2018, April 24). Trump Voters Driven by Fear of Losing Status, Not Economic Anxiety, Study Finds. *The New York Times*. Retrieved from www.nytimes.com

[19] We have considered a different task: predicting Trump's approval rate dynamics, since we can model this at the state level rather than at the district-level. which can produce reasonable results. See GitHub for the relevant code.

Appendix

A1. Psychometric Weighting Schemes

To illustrate each weighing scheme, we focus on the climate change topic, which attempts to measure a respondent's level of climate change skepticism. In particular, we use the following survey question: "How do you feel about government spending and regulations to address climate change?"

Answer Choice	Weighting Scheme #1	Weighting Scheme #2	Weighting Scheme #3
"Favor Strongly"	0	0	0
"Favor Weakly"	0	0	1
"Don't know"	0	0	2
"Oppose Weakly"	1	1	3
"Oppose Strongly"	1	1	4
Domain of α_{topic}	0 or 1	0 to 7	0 to 24

Table 4: How weights are assigned to different answer choices and the range of integer values that α_{topic} can take under each weighting scheme.

Under weighting scheme #1, we assign 1 point for answers that indicate skepticism towards climate change and 0 points otherwise. We assign $\alpha_{topic} = 1$ if the respondent receives one point for at least half of the questions, and $\alpha_{topic} = 0$ otherwise. A value of 1 indicates that the respondent appears to identify with the trait that we are interested in. In the above example, $\alpha_{climate} = 1$ means that a respondent appears to be skeptical about climate change.

Under weighting scheme #2, we assign points to answer choices in the same manner as in weighting scheme #2. However, instead of assigning $\alpha_{topic} = 0$ or $\alpha_{topic} = 1$ based on a threshold, we assign the sum of the points that the respondent received from all of the questions. This allows us to represent how strongly the respondent identifies with a trait using an integer scale (from 0 to the number of questions for that topic) rather than a binary label. In the case of the climate change topic, we would rate respondents on a scale of 0-7 since there are seven questions for that topic.

Under weighting scheme #3, we assign a varying number of points to each answer choice depending on how much that answer reflects the trait that we are interested in. In the case of the climate change topic, we would rate respondents on a scale of 0-24, with 24 meaning that the respondent chose the most extreme answer for every question. This scheme allows us to consider how strongly the respondent identifies with a trait on a per-question basis. However, there is a risk of adding noise to the model if people do not actually identify with these traits in such a fine-grained manner. For example, it may be the case that there is no meaningful difference in the level of climate change skepticism between a respondent with $\alpha_{climate} = 23$ and one with $\alpha_{climate} = 24$.