The Elusive Likely Voter: Improving Electoral Predictions with More Informed Vote Propensity Models

Anthony Rentsch
Harvard University
anr222@g.harvard.edu

Brian F. Schaffner
Tufts University
brian.schaffner@tufts.edu

Justin H. Gross
University of Massachusetts Amherst
jhgross@polsci.umass.edu

January 22, 2019

Abstract

Political commentators have offered evidence that the "polling misses" of 2016 were caused by a number of factors. This project focuses on one explanation, that likely voter models – tools used by pre-election pollsters to predict which survey respondents are most likely to make up the electorate and, thus, whose responses should be used to calculate election predictions – were flawed. While models employed by different pollsters vary widely, it is difficult to systematically study them because they are often considered part of pollsters' methodological black box. In this paper, we use Cooperative Congressional Election Study surveys since 2008 to build a probabilistic likely voter model that not only takes into account the stated intentions of respondents to vote, but also other demographic variables that are consistently strong predictors of both turnout and over-reporting. Using this model, which we term the Perry-Gallup and demographics (PGaD) approach, we show that we are able to reduce the bias and error created by likely voters model to a negligible amount. This likely voter approach uses variables that pollsters already collect for weighting purposes and thus should be relatively easy to implement in future elections.

Funding

Funding for this work was provided by the National Science Foundation [grant numbers 1225750, 1430505, 1559125].

Introduction

Forecasting who will actually vote is a particularly challenging problem because, unlike other estimation efforts, pollsters must make an inference about a population (all eventual voters) that does not yet exist. So, unlike the standard sampling problem of identifying an existing population, selecting a suitable sampling frame that comes close to the true list of all population members, and then conducting some variant of probability sampling, here, we are dealing with two distinct, but not independent, stages of estimation. Pollsters must identify the subset of citizens who will vote even though people do not correctly identify themselves as members of this subset. Added to this challenge is the fact that pollsters often treat their likely voter models as proprietary, which makes it difficult for scholars to adjudicate amongst these models. For example, the American Association of Public Opinion Researchers' (AAPOR) 2016 election post-mortem report noted that likely voter models may have contributed to polling errors, but due to the lack of available data the authors were unable to conduct a full exploration into this factor (2017).

In this paper, we build on previous work (e.g. Keeter, Igielnik, and Weisel 2016) to develop a framework for incorporating indicators of the likelihood of voting that builds on political science scholarship and will be reasonably straightforward for pollsters to implement in future elections, using information that they typically already collect. We leverage the Cooperative Congressional Election Study (CCES) surveys taken during presidential and midterm election years over the last decade to build and evaluate different likely voter models. Using this data, we consider different approaches to the identification of likely voters and ultimately argue against the common practice of hard classification of likely vs. unlikely voters and in favor of weighting eligible voters' potential vote choice by an estimated probability that they will vote, using a combination of respondent self-reported intention to vote, voting history, and demographic data pollsters already collect. We show that this approach, which we call PGaD (Perry-Gallup and Demographics) performs best at predicting which respondents actually vote and when it comes to vote share estimates.

The Unique Problem of Likely Voters

Pollsters face a unique challenge when it comes to the task of making inferences about an electorate that has not yet formed. The typical approach that a survey firm takes is to sample from some larger population (e.g. adults or registered voters) and then attempt to identify the subgroup from among that set of individuals who will actually vote. In practice, attempts to estimate a result for likely voters layers the problem of measurement error on top of the challenges of sampling error that pollsters already face. This is because pollsters are attempting to make an inference about a subgroup of the population that (1) often has different

Table 1: Intention to vote in 2016 CCES and validated turnout

Do you intend to vote in the 2016 general election?	% of group voting
Yes, definitely (78%)	64%
I already voted (3%)	68%
Probably (7%)	29%
Undecided (5%)	18%
No (7%)	9%

Note: N = 64,600 respondents to the 2016 CCES. Percentages calculated using post-stratification sampling weights.

preferences than the larger population and (2) does not accurately identify itself when asked.

To illustrate the second point, Table 1 shows responses to a question on the 2016 CCES asking respondents whether they intend to vote in the upcoming general election. About four-in-five respondents report that they have either already voted (early) or that they will definitely vote. Yet, only about two-thirds of this group were matched to a valid vote record. In essence, there is a high rate of mis-identification for this question meaning that pollsters cannot easily accurately identify the subgroup of voters. Indeed, if one-third of respondents misreported their sex, we would have little confidence in using our surveys to make inferences about the sub-populations of men or women.

The people most likely to over-report voting (to say they will vote when they will not) are "those who are under the most pressure to vote" (Bernstein, Chadha, and Montjoy 2001). For example, Ansolabehere and Hersh (2012) find that "well-educated, high-income partisans who are engaged in public affairs, attend church regularly, and have lived in the community for a while" retrospectively misreport their voting behavior. Other studies suggest that young, black, and nonpartisan respondents tend to misreport (Rogers and Aida 2014), while individuals from older age brackets and who are highly educated are more likely to accurately assess their voting likelihood (Pew Research Center 2000).

Misreporting poses a problem for pollsters because it introduces systematic error – people who vote more often are demographically and politically different than people who vote less frequently (Keeter, Igielnik, and Weisel 2016), despite a lack of substantial differences between self-reported voters and self-reported nonvoters (Rogers and Aida 2014). Since many of the variables tied to misreporting are also associated with partisanship, polls that do not report their results in terms of likely voters generally overestimate support for Democrats (Newport 2000). Thus, using self-reported voting intention can produce biased estimates, which is a problem when one out of every four nonvoters reports that they will vote (Freedman and Goldstein 1996). Returning to our example, if we look at people who said they would definitely vote or already voted in the 2016 pre-election wave of the CCES, Clinton held a 7.6-point margin over Trump. However, among those who eventually had a record of voting, her margin was just 4.5 points. Thus, the issue that pollsters face is clear. Respondents do not accurately indicate whether they will actually become voters, and this misreporting is not independent of candidate preference.[1]

Existing Approaches to Identifying Likely Voters

To account for misreporting, most likely voter models include a combination of questions about an individual's vote intent, voting history, and interest in politics. The types of questions used vary; sometimes they are as simple as asking respondents if they plan to vote. Some pollsters construct composite indices based on a series of questions (Keeter et al. 2016). Sometimes these indices focus exclusively on the combination of past voting behavior and vote intent (Freedman and Goldstein 1996; Murray, Riley, and Scime 2009), while others include questions about the voting process, such as where the respondent's polling place is (Kiley and Dimock 2009). One prominent example is the Perry-Gallup index, which is composed of seven questions that capture how much thought a respondent has given the upcoming election, whether the respondent has ever voted in their current district, how closely the respondent follows government and public affairs, how often the respondent votes, the respondent's vote intent for the upcoming election (there are two questions on this), and the respondent's vote history (Keeter et al. 2016). Respondents are assigned points based on their responses to those questions and those who achieve a certain number of points are considered likely voters.

Once pollsters have settled on a particular set of questions that they use to determine likely voters, they then have to decide how to use those questions to make an inference about likely voters. Deterministic likely voter models create a likely voter score for each respondent and then create a decision rule to decide whether to include or exclude a response when calculating election predictions. As such, this approach is often referred to as a cutoff approach or threshold model, as pollsters decide which responses to consider and which ones to discard for their final predictions.

On one hand, simplicity is a virtue of deterministic models. Some respondents end up voting and some will not, so it makes sense to model this reality by including the responses of those who are most likely to vote and excluding the responses those who are least likely to vote. Deterministic models are simple to justify and explain to a broad audience because they resemble how an election works. On the other hand, cutoff approaches suffer from the loss of information that comes from categorizing all respondents as either voters or non-voters. Probabilistic models, which are rarely implemented by political polling firms, offer some clear benefits. In a probabilistic approach, each respondent is assigned an estimated probability that they will vote. This probability is then used as a weight: responses from those who are more likely to vote are weighted more heavily than responses from those who are unlikely to vote, but all are included in

[1] We use the 2016 CCES for illustrative purposes, but similar patterns can be found in other years and in other surveys.

the election prediction. Pollsters who employ such a strategy generally use the same predictors featured in deterministic models to predict validated turnout in surveys from previous election cycles and then apply the model to data from the current cycle to generate a predicted probability that each respondent will vote in the upcoming election (Keeter et al. 2016).

One benefit of probabilistic models is that they allow pollsters to consider more information; the preferences of all respondents are still considered, just to an extent proportional to their assessed probability of actually voting. Related to this advantage is the fact that a probabilistic approach is a principled way to actually estimate the behavior of the yet-to-be-determined population of interest: eventual voters. Put simply, if we want to estimate the proportion of voters who will vote for Candidate X, we should be ascertaining the probability that each eligible voter will vote for Candidate X and then take the mean over all voters. But the probability of voting for Candidate X is really a joint probability of voting and specifically voting for that candidate. Letting $V =$ the event a person votes and $X =$ the event of preferring candidate X, and applying the general product (or chain) rule, a basic probabilistic principle, we have $P(V, X) = P(X|V)P(V)$. That is, the joint probability a person will vote and cast that vote for Candidate X is simply the probability of casting a vote for Candidate X given they vote at all, multiplied by the probability that they vote. Since we can't estimate this for every eligible voter, we sample eligible voters and apply post-stratification sample weights in the usual way. From this perspective, the estimated probability of voting is not just another kind of weight to be added to sampling weights; instead it is part of a simple model for the yet unseen population.

A Demographics-Informed Probabilistic Model of Likely Voters

We see the probabilistic approach as superior to deterministic models, for the reasons outlined above. However, we also note that once a survey researcher has decided to use this approach, they need not limit themselves to the set of questions that are typically used in deterministic models. In particular, we note that pollsters typically ask respondents a number of questions for the purpose of post-stratification weighting that are also powerful predictors of turnout. A vast literature demonstrates the primacy of socioeconomic status (education in particular) and age in differentiating voters from nonvoters (Verba and Nie 1972; Wolfinger and Rosenstone 1980; Blais 2006; Leighley and Nagler 2013). Political activism, ideological extremism, and race are all tied to the likelihood to vote as well (Verba and Nie 1972).

Fortunately, many of these variables that scholars have identified as important predictors of turnout also happen to be regularly collected in surveys, typically for the purpose of post-stratification weighting. Thus, rather than leave these factors on the table, we propose that pollsters incorporate them into their likely voter models. Specifically, we propose a probabilistic likely voter model that not only uses the typical items (such as those from the Perry-Gallup index), but also adds demographic information about respondents such as age, education, race, income, gender, and strength of partisanship. Since we know that these variables are consistently related to turnout, they should also help us produce more precise likely voter probabilities. Indeed, as we will show, this approach provides a substantial improvement on modeling likely voters when it is applied to the 2014 and 2016 CCES surveys.

Design and Modeling Approaches

In order to build and assess the performance of likely voter models, we use CCES surveys fielded during presidential (2016, 2012, and 2008) and midterm (2014 and 2010) election over the past decade. There are several reasons why the CCES is a desirable data source for the task of modeling likely voters. First, since 2008 the CCES has used vote validation, which is considered the gold standard for studying individual-level election turnout. After the election, CCES respondents are matched into Catalist's national voter file database, which consists of over 240 million unique voting-age individuals across the United States that the organization has compiled by collecting voter registration records from each state and combining these records with commercial records purchased from data aggregators. Their database allows their clients, such as the CCES, to identify with a high level of accuracy which individuals have a record of voting in a certain election and which individuals do not (Ansolabehere and Hersh 2012). A validated vote record for each respondent is the key dependent variable for the probabilistic likely voter models, which require that a relationship between a number of covariates and turnout be estimated from historical data and then applied make predictions from new data.

Second, the CCES is a high quality nationally representative large-N survey, which allows for generalizability of the sample not only to the national population, but also to the populations in each state. The state with the fewest total respondents over the five years we consider (Wyoming) still has nearly 500 valid observations across the several elections. Most states have thousands of observations across these elections.

Finally, the CCES asks a wide range of demographic and attitudinal questions that may be useful for identifying likely voters based on the literature about misreporting and turnout. The survey items we consider are vote intention, vote choice, self-reported previous turnout history, voter registration status (self-reported), interest in politics, age, gender, education, race, income, and partisanship.

There are, however, a few limitations to the CCES data. First, there is a small amount of inconsistency in the operationalization of the vote history variable. For 2012 and 2016, the indicator variable is coded 1 if a respondent reported that they voted in the previous presidential election and 0 if they were unsure, did not recall, or reported that they did not vote. In 2010 and 2014, it is coded similarly, using a respondent's vote history in the previous presidential election rather than the previous midterm election. In 2008, respondents were not asked about their voting behavior in the 2004 presidential election, so whether or not they reported that they voted in a primary election or caucus in 2008 is used as a proxy. Although this is not a perfect substitute, there is substantial correlation between the two questions in the 2016 sample; in 2016, over 72 percent of respondents who voted in the 2012 presidential election also voted in the 2016 primary, while over 78 percent of those who did not vote in 2012 also did not vote in a primary or caucus in 2016.

Second, the state of Virginia did not release turnout records to Catalist for the 2008 and 2010 election cycles. Thus, we remove Virginia respondents from our dataset for those two years. Between the five CCES surveys we consider, there are 263,535 observations. After removing mis-coded and missing values, our final data set contains 259,940 observations across five election cycles.

We compare several approaches to defining likely voters in our analysis. For each approach, we evaluate how well the models predict individual-level turnout and how well the models produce accurate estimates of the vote margin for president in 2016 and for the House of Representatives in the case of our analysis of the 2014 midterm elections. For the purposes of predicting 2016 turnout, we use data from 2008, 2010, 2012, and 2014 as the training data. For predicting 2014 turnout, we use 2008, 2010, and 2012 as the training data.

When we examine how well each likely voter approach approximates the vote margin, we operationalize that margin as the difference between the percentage of validated voters in the CCES pre-election poll who said they intended to vote for the Democratic candidate minus the percentage who intended to vote Republican. We use this as a benchmark rather than the actual election outcomes because our interest is in examining how well we can approximate the electorate's preferences as they stood when the poll was conducted. Last minute shifts in vote preferences or faulty post-stratification weighting might cause a poll to miss the final election outcome even if that poll perfectly predicted who would vote. Since our interest is in isolating which likely voter model is most effective, we want to eliminate these other sources of error from consideration and focusing on the vote share among the respondents in our survey who actually voted is the best way to do this.

We consider four approaches to modeling likely voters. In the first and simplest approach, we compare responses to the vote intention question to a threshold value in order to identify likely voters. Specifically, we identify as likely voters everyone who says either that they have already voted (early or absentee) or that they will definitely vote. While we could extend our threshold to include those who say that they will "probably" vote, doing so provides a much less accurate prediction of our quantities of interest than limiting the pool of likely voters to those in the "already voted" and "definitely" categories.[2]

[2] See the Supplementary Material for results using other thresholds.

The second cutoff approach is a reformulation of Pew's Perry-Gallup index. The CCES does not contain all of the questions that are used in the Perry-Gallup index and the question wording varies for the items that do appear. As an approximation of the Perry-Gallup approach, we use vote intent, vote history, and political interest from the CCES to create a version of this index. Together these questions capture five of the seven items on the Perry-Gallup index; the one dimension they do not capture is self-reported historical voting behavior, as the CCES does not ask about whether a respondent has voted in their district before or about their voting frequency. We assign respondents points based on the follow criteria: two points for those who reported that they already voted (early or absentee) in the 2016 general election and those who report they will "definitely" vote, and one point for those who will "probably" vote in the election. Respondents who reported that they voted in the 2012 general election are awarded one point. Those who follow what is going on in government and public affairs "most of the time" are given two additional points, while those who follow "some of the time" are awarded one additional point. We make two further adjustments. We give respondents who report that they are registered to vote one point. Further, since respondents who are younger than 22 would not have had the chance to vote in the previous election, they are given one additional point. The minimum score in this version of the Perry-Gallup index is zero while the maximum score is six. We create likely voter subsets based on these scores and compute the accuracy of our predictions. In the paper, we focus on two different cutoffs – just taking those who score a 6 and also taking those who score a 6 or a 5. These two groups provide the most accurate estimates of turnout and vote margins among the possible cutoff points (but see the Supplementary Material for results from all possible cutoffs).

The final two models that we estimate are the two probabilistic likely voter models. Both of these models employ random forests, a powerful machine learning tool that uses a large number of decision trees, each fed with a random subset of the data and a random subset of all possible variables at each split, that can be used to compute vote propensity scores much in the same way that logistic regression can be used.[3] The benefit of random forest algorithms is that, since they randomly sample the data and the available predictor variables, they avoid much of the bias that traditional decision tree approaches encounter, which make them especially useful for prediction. The random forest algorithm outputs predicted class probabilities for each respondent; that is, each respondent is assessed a probability that they will vote. We then weight each respondent according to that probability (as well as according to their post-stratification sampling weight). Thus, an individual who received a vote propensity score of .2 would contribute only-quarter as much influence to the likely voter estimates as an individual with a vote propensity score of .8.

The first probabilistic random forest model that we produce uses the variables that comprise the modified

[3] These approaches can also be implemented with logistic regression and doing so produces similar results, as we show in the Supplementary Material.

Perry-Gallup index that we described above – intention to vote, previous turnout, registration, interest in politics, and whether the person was eligible to vote in the previous election. The second model includes all of these items, but adds a set of common demographic variables that most pollsters routinely collect for the purposes of weighting and analysis. These variables include age, race, education, gender, family income, and partisan strength.

For the national validation, we pool all of the observations together as if they were fielded as a part of a national poll. The models' performance is evaluated using the full national sample in 2016 and the random forest models are trained using all of the data from the previous CCES surveys. We also estimate state models. In that test, we evaluate each type of likely voter model 51 times (once for each state plus the District of Columbia) using only data from the given state. Further, we train each state's model only with historical data from that state to isolate unique characteristics of each state. For instance, the likely voter model that is applied to 2016 CCES respondents from Texas is only trained on data from other respondents from Texas in earlier cycles.

Results

We begin by summarizing how each of the various likely voter approaches fares in predicting the 2016 presidential vote preference, both nationally and by state. Table 2 presents this information, in aggregate in the case of the states. The first column shown for each model is what we call *implied turnout*. This is simply the percentage of the weighted sample that each model implies will be voting.[4] For the cut-off models, this is a straightforward calculation indicating the share of respondents who meet a particular criterion. For example, 70.78% of the 2016 CCES respondents indicated that they had already voted or would definitely vote. For the probabilistic models, we take the average of the turnout propensity scores to calculate the implied turnout among the sample. As a baseline, 55.11% of the 2016 CCES respondents were validated voters.

The second and third column entries are for the Democratic bias produced by each likely voter model. Democratic bias is simply the difference between the margin by which a set of likely voters preferred Clinton over Trump and Clinton's actual margin over Trump among all validated voters in the sample. Recall that we use Clinton's margin over Trump among validated voters in the sample as our baseline (rather than the actual election outcome) in order to isolate the effect of different likely voter models on producing bias. Using the margin among validated voters in the sample allows us to eliminate other reasons that the survey may have produced an inaccurate result, such as sample composition or systematic changes in vote preferences after

[4] All of these calculations are based on using the post-stratification weights provided with the survey to ensure the sample is nationally representative.

Table 2: Democratic bias in survey estimates of Presidential vote preference based on different approaches to defining likely voters, 2016.

Approach	Implied Turnout	National Bias	Avg. Bias by State	Avg. Absolute Error by State
		Cutoff Approaches		
Already voted + will definitely vote	70.78%	3.59	2.46	4.44
Perry Gallup 6's	41.66%	-1.70	-2.76	6.19
Perry Gallup 6's + 5's	60.26%	2.05	1.18	3.98
		Probabilistic Approaches		
Perry Gallup	66.55%	3.29	1.75	4.17
Perry Gallup + Demographics	59.86%	-0.19	-0.36	4.02

the survey was conducted. Among validated voters in the 2016 CCES, Clinton enjoyed a 2.92 percentage point advantage over Trump, not too far from her actual margin of 2.1 percentage points. The final column calculates the average absolute error for the state-by-state modeling.

Basic Cutoff Approaches

The most basic approach to determining likely voters is simply to ask survey respondents if they intend to vote and then identify as voters those who say that they will definitely vote or, in a less restrictive case, those who will probably or definitely vote. In Table 2 we show the results from identifying only those who said that they had already voted or will definitely vote as likely voters. Even with this most restrictive definition, about 71% of CCES respondents would be likely voters. Additionally, this simple determination of likely voters leads to over-stating Clinton's national margin by 3.59 percentage points. This is a fairly large bias – one that would mean the difference between a very close contest and a likely landslide.

The Perry-Gallup method is a somewhat more involved approach to defining a cutoff value for determining who should qualify as a likely voter. In Table 2 we show two possible definitions of likely voters from our modified Perry-Gallup scale – one that is fairly restrictive and one that is less so. The most restrictive Perry-Gallup cutoff uses just those scoring the maximum score of 6 on the items. Under this approach, just 41.66% of CCES respondents are identified as likely voters. With this restrictive approach, Clinton's margin is actually estimated to be 1.70 percentage points lower than it actually was. The more inclusive approach (taking those who score either a 5 or a 6) leads to 60.26% of respondents being considered likely voters. With this group of likely voters, Clinton's margin is predicted to be 2.05 percentage points larger than it was among validated voters in the survey.

The final two columns in Table 2 show the average Democratic bias and the average absolute error across the states when we generate likely voter groups state-by-state. For the first two cutoff models, the average bias exceeds two percentage points (the first favoring Democrats and the second favoring Republicans), while it is just 1.18 points for the more inclusive version of the Perry-Gallup cutoff. The average absolute error is also smallest for the more inclusive Perry-Gallup cutoff model. Among all cutoff approaches, the Perry-Gallup model works best, though for a national estimate the more restrictive cutoff appears in this case superior while for the state-by-state analysis the less restrictive approach produces less bias/error.

Probabilistic Models

Now, we turn to examining a probabilistic approach to defining likely voters. We show results from two different types of probabilistic models – one in which we simply use the Perry Gallup variables to predict turnout and another in which we also include demographic predictors. We use the random forest approach to predict validated turnout in CCES surveys from previous years, and then use the model trained on those earlier surveys to estimate each respondent's probability of voting in the test year. These vote propensity scores are then used as an additional weight to calculate the vote margin between Clinton and Trump.

Returning to Table 2, we see that the Perry-Gallup probabilistic approach does not improve upon the Perry-Gallup cutoffs. Indeed, using the Perry-Gallup items in a probabilistic model yields an implied turnout of 66.61% and produces a pro-Democratic bias of more than 3 percentage points. Somewhat surprisingly, this is significantly worse than either of the Perry-Gallup cutoff models.

Finally, we show the results from our proposed approach, which uses the Perry-Gallup items plus a variety of demographic variables (PGaD) to generate a vote propensity score for each respondent. Table 2 demonstrates the value of this approach. First, this model produces an implied turnout of 59.81%, which is reasonably close to the actual turnout of 55.11% among 2016 CCES respondents. More importantly, the bias produced by this approach is negligible. Indeed, the national vote share from the Perry-Gallup plus demographics model is just two-tenths of a percentage point off the actual margin among voters in the 2016 CCES. In the state-by-state analysis, the PGaD model's average bias is just -0.36 and the average absolute error across the states is 4.02 points.

Evaluating the Probabilistic Models

Figure 1 evaluates how well each of the probabilistic models fares at predicting actual turnout by plotting respondents' vote propensity scores against actual turnout rates. An extremely accurate model would produce a line that fell exactly along the dashed 45 degree diagonal line. The left-side plot shows why only using the Perry-Gallup items fails to produce particularly accurate results. This model struggles to sort voters from non-voters in the middle of the vote propensity scale. In fact, turnout rates were actually higher among respondents receiving a vote propensity score of 40 than they were for those receiving a propensity

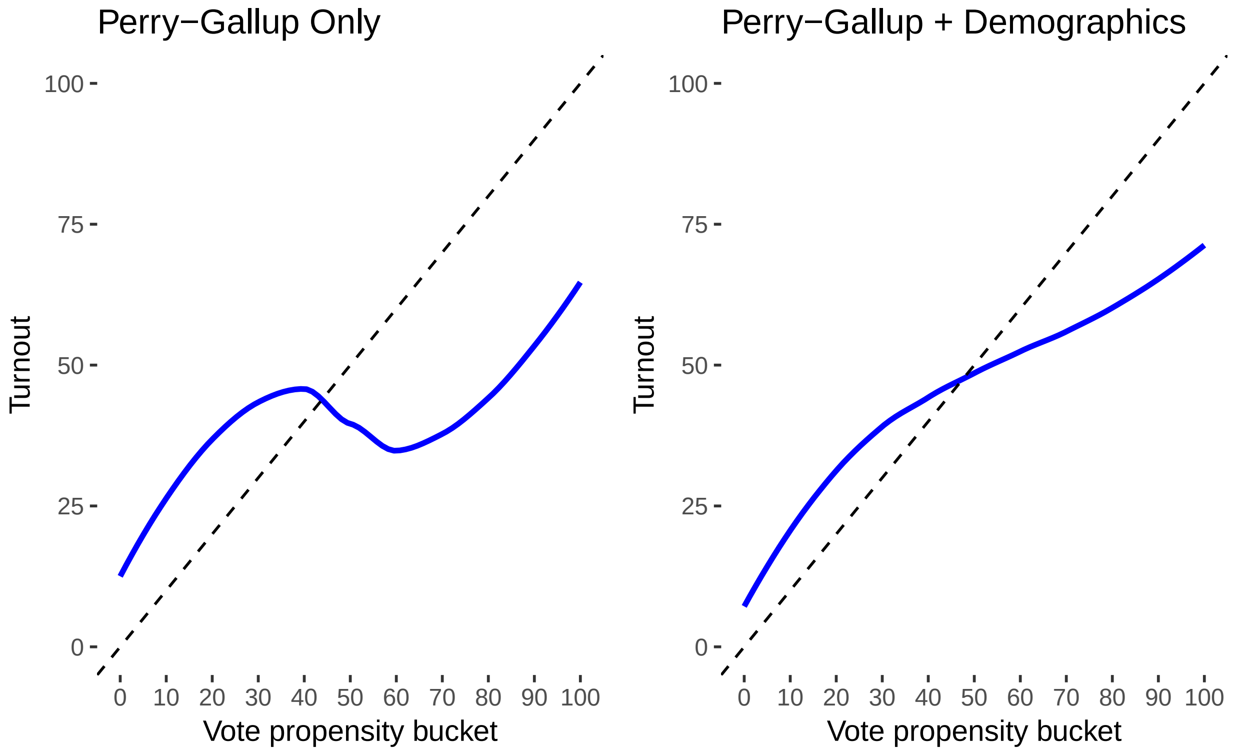

Figure 1: Validated turnout rate based on respondent's turnout propensity score for 2016 CCES test. Propensity scores generated using random forest models.

score of 60. The model is also under-performing among those to whom it assigns a high probability of voting, perhaps due to the fact that people tend to over-report on many of the key Perry Gallup items.

The blue line for the Perry-Gallup plus demographics model tracks much closer to the 45-degree diagonal. Overall, the model slightly under-predicts turnout among those assigned a propensity score of 40 or lower and it slightly over-predicts for those at 50 or above.

Figure 2 provides additional insight into how adding demographic information helps to predict likely voters. The figure simply shows the distribution of the predicted vote propensity scores for 2016 CCES respondents generated by each model. When we use only the Perry-Gallup items, most of the propensity scores are clustered either at 0 or 1, with very few respondents receiving marginal propensity scores. However, adding demographic information to the model provides much more nuance. Indeed, the number of respondents receiving a probability of 1 drops by half and the cases become much more distributed across the spectrum of propensities. Overall, the Perry-Gallup items alone are a fairly blunt instrument that fails to sort voters meaningfully with regard to vote propensities, a problem that is much improved upon once we add the additional demographic items.

But which items are most influential in helping us to identify likely voters? One of the useful metrics

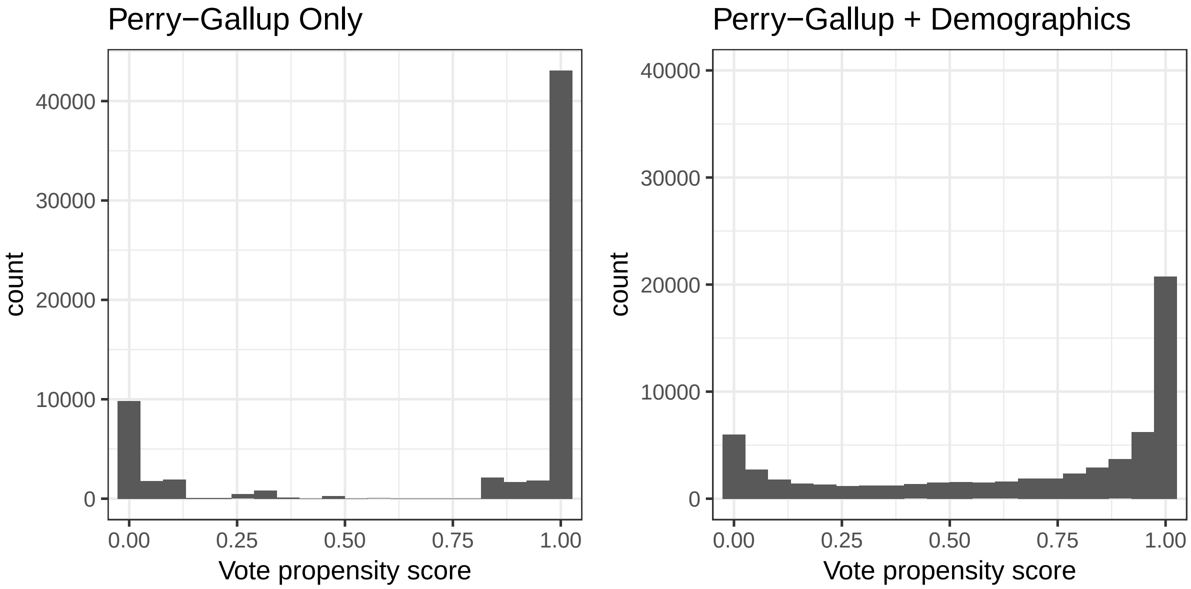

Figure 2: Distribution of Propensity Scores Using Each Probabilistic Model, 2016.

produced by a random forest is the importance of each variable for predicting the outcome. Here we use the Mean Decrease in Accuracy metric. The basic logic of this measure is to see how much less predictive the random forest is at correctly classifying observations when each variable is randomly perturbed. If the predictive accuracy of the model drops considerably when a variable's values are randomly assigned, then that variable is deemed to be more important for predicting the dependent variable.

Figure 3 shows variable importance plots for both random forest models that we estimated for the 2008, 2010, 2012, and 2014 datasets. The left-side plot shows the importance of each of the Perry-Gallup items we used in our simple model. The plot on the right includes this set of variables as well as the demographic items from our more complex model. Larger values on the x-axis indicate that the variable is more predictive of being a validated voter.

Note that a respondent's stated intention to vote is the most important predictor in both models. As we noted before, this is largely because the question is a very good predictor of non-voting; if somebody says that they do not intend to vote, it is very likely the case that they will be a non-voter. However, the vote intention question is less powerful as a predictor of voting, since many people who say that they will definitely vote ultimately fail to do so. Among the Perry-Gallup items, registration status is the next most predictive variable, followed by interest in politics, and then turnout history.

However, recall that the Perry-Gallup items alone fail to sufficiently distinguish between voters and non-voters, producing a fairly large pro-Democratic bias in 2016. By contrast, adding demographics to the model reduced the bias considerably and brought the turnout rate among the sample down closer to what it actually

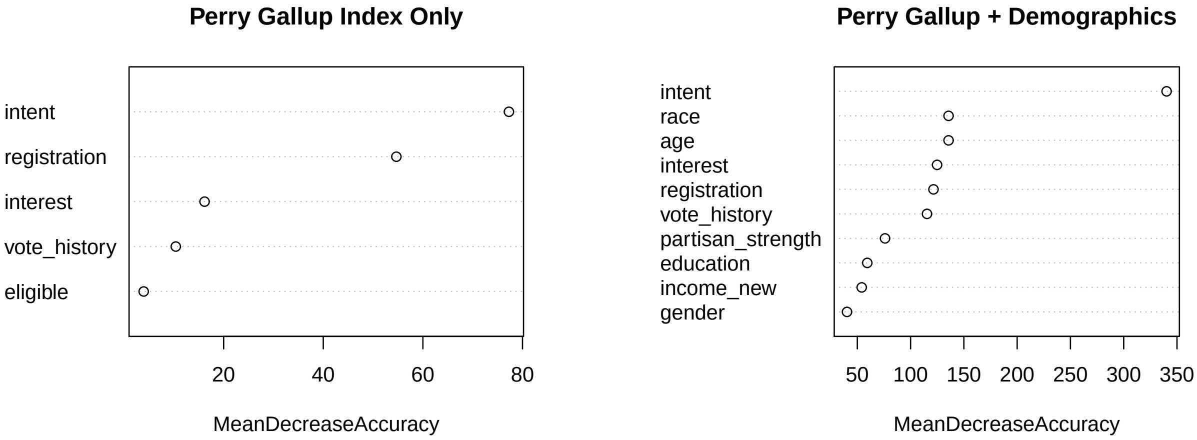

Figure 3: Variable Importance Plots showing importance of each item to predicting validated turnout, 2016

was for the 2016 CCES respondents. The right-side plot in Figure 3 shows that among all of the additional demographic items, *race* and *age* were the most important for helping to discern voters from non-voters (and those two variables were second only to vote intention in terms of overall importance).

Why does adding demographic variables produce such an improvement in the probabilistic model? As we noted before, the vote intention question is quite useful at distinguishing non-voters – just 9% of individuals who say they do not plan to vote nonetheless show up as validated voters. Unfortunately, other responses to the vote intent question are not nearly as predictive. Only 29% of people who report that they will probably vote actually do so, and not even two-in-three who say they will definitely vote actually turn out. This is where the demographics help the model – by separating out people whose reported intention to vote is a stronger (or weaker) signal about what they will actually do.

Figure 4 demonstrates the value of adding a variable such as *age* to the model. In particular, age is strongly predictive of turnout among two key groups – those who say that they will definitely vote and those who say that they will probably vote. For example, among respondents who were 70 years old and who said that they would definitely vote, 80% actually voted. However, among respondents who were 20 years old and said they would definitely vote, turnout was just about 50%.

The role of age is even more clear when one compares percent voting as a function of age for probably, undecided, and no. Among 20 year-old respondents, there is no difference in turnout rates among those saying that they do not plan to vote and those saying that they are undecided. Yet, among 60 year-old respondents, the undecided group is twice as likely to vote as those saying no. Meanwhile, among 20 year-old respondents, respondents saying that they will probably vote are about twice as likely to vote as those saying that they will not vote. Yet, among 60 year-old respondents the “probably” group is *four* times more

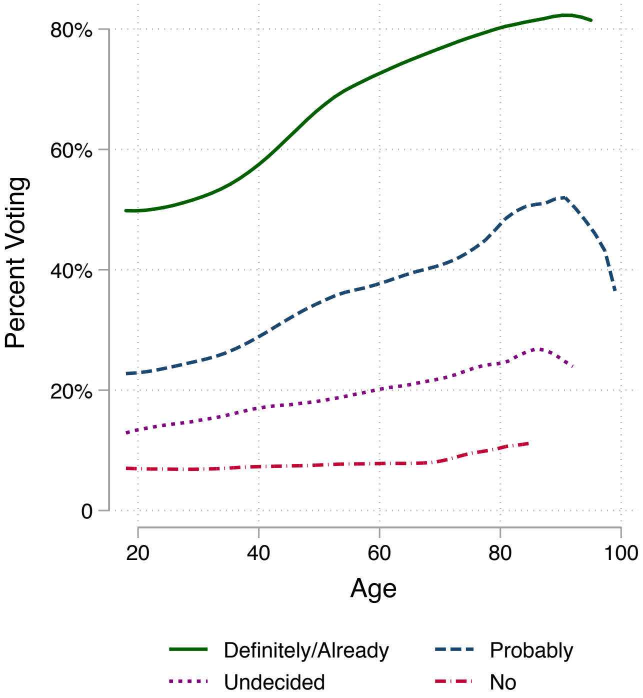

Figure 4: Actual Turnout of 2016 CCES Respondents Based on Age and Intention

likely to vote than the no group.

Extending the Model to a Midterm Election

So far, we have demonstrated that our probabilistic Perry-Gallup plus demographics (PGaD) approach produces accurate projections about turnout and the presidential vote margin in 2016. However, a reasonable question is whether this approach would work equally well in a non-presidential election year. To investigate, we re-produced the analysis from above, but this time focusing on predictions for the 2014 CCES midterm election survey. For this validation task, we trained our models only on the 2008, 2010, and 2012 CCES surveys and then used those models to make prediction about turnout in 2014. Our measure of bias for this analysis comes in the form of the national vote for the House of Representatives. In our 2014 CCES sample, 50.1% of respondents were validated voters and they preferred Republican House candidates over Democrats by a margin of 5.57 percentage points.

Table 3 shows the implied turnout and national Democratic bias estimates for each approach to defining likely voters. In almost every case, the implied turnout rates are much higher than the actual percentage of respondents who were validated as voters. Part of the reason for this discrepancy may be due to the fact that our model is trained on data from one (relatively high-turnout) midterm and two presidential elections,

Approach	Implied Turnout	National Bias
	Cutoff Approaches	
Already voted + will definitely vote	73.01%	2.41
Perry Gallup 6's	44.26%	-5.06
Perry Gallup 6's + 5's	66.35%	1.20
	Probabilistic Approaches	
Perry Gallup	75.11%	2.42
Perry Gallup + Demographics	69.89%	0.50

Table 3: Democratic bias in survey estimates of House vote based on different approaches to defining likely voters, 2014.

whereas turnout in 2014 was the lowest turnout rate of any federal election since 1942.[5] Nevertheless, the Perry Gallup plus demographics model again produces the least amount of bias in predicting the national House vote margin compared to the other approaches. In this case, the approach produces just a half-percentage point pro-Democratic bias in terms of predicting the House vote margin. The next most accurate approach is the Perry-Gallup cutoff that includes respondents scoring 5 points or higher. With this approach, the pro-Democratic bias was 1.20 percentage points.

In both 2014 and 2016, the probabilistic, demographics-augmented Perry Gallup model does the best, keeping bias well below one percent.

Simulations

As a final method of comparing these approaches, we sought to examine how each model would perform on sample sizes that pollsters more typically deal with and across multiple samples. To do this, we took 1,000 random samples each from the 2016 CCES data with sample sizes of 500, 800, 1,000, and 1,200. For each sample, we calculated the Clinton vote margin produced by each of the five models. Figure 5 shows the distribution of vote margin estimates for each of these approaches at each of the four different sample sizes. Notably, the plots show that the PGaD model provides the best combination of unbiased estimates paired with less variability. The cutoff method of taking only those scoring 6's on the Perry Gallup scale has, on average, low bias, but a wider distribution. This means that on average that approach will produce more error. For example, among the simulated samples with N = 1,000, the PGaD method produced a vote share margin that was within 3 points of the true margin 43% of the time, while the Perry-Gallup cutoff approach did so just 34% of the time. Similarly, the average absolute error of the PGaD approach is at least 0.75 points lower than the Perry Gallup cutoff (and all other methods) at each sample size. More detailed statistics from these simulations are available in the Supplementary Material.

[5] `http://www.electproject.org/national-1789-present`

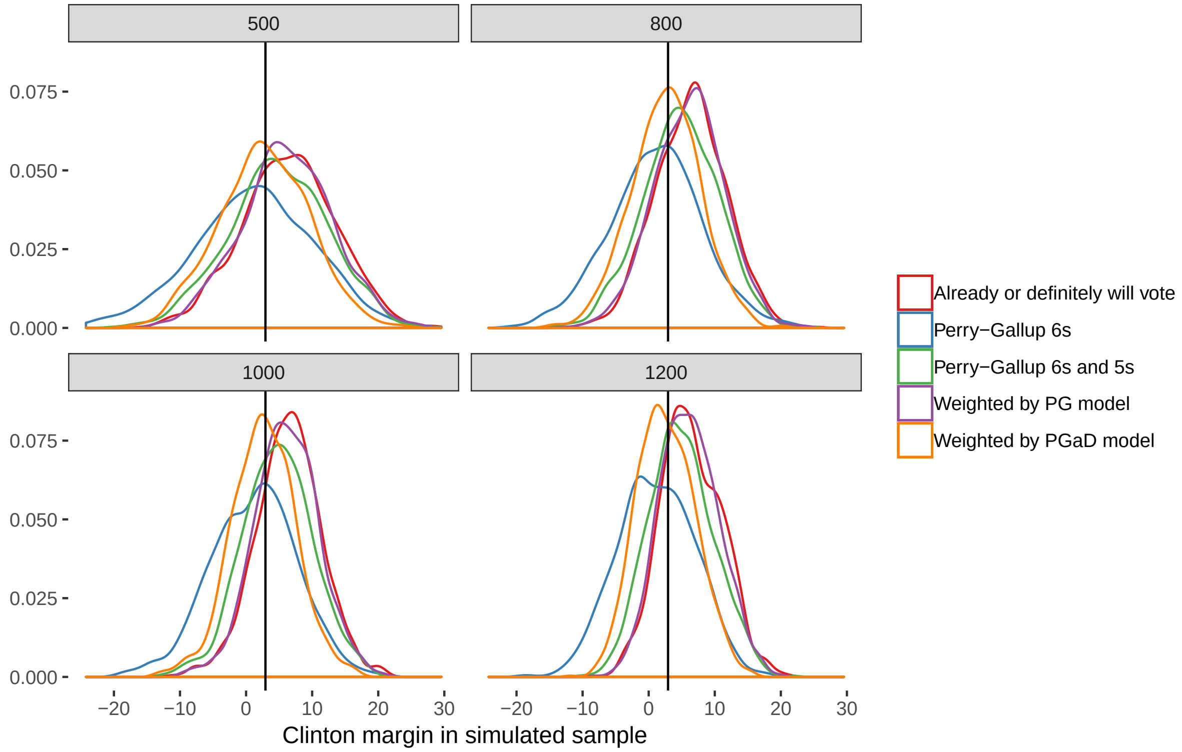

Figure 5: Distribution of vote margin estimates of simulated samples from 2016 CCES

Conclusion

In this paper, we have compared a number of common approaches to the likely voter problem, including three examples of predominant cutoff (or threshold) techniques and two probabilistic models. We found that the most successful Perry-Gallup-style threshold-based classification of likely voters led to a Democratic estimation bias for the national vote of 1.20% for the midterm election (2014) and 2.05% for the presidential election (2016). Among the probabilistic models, our Perry-Gallup + Demographics approach (PGaD) beats all the other estimates. The absolute bias (or aggregate error) is no more than a half-percent for each election (nationwide Democratic bias of 0.50% in 2016; national Republican bias 0.19% and average Republican state bias 0.36% in 2016).

Based on these results, we encourage pollsters to consider implementing likely voter models (such as PGaD) that maximize the full amount of information at their disposal. Specifically, a likely voter model that is probabilistic uses information from all respondents in the sample rather than discarding those that fail to meet a particular threshold. And a likely voter model that makes use of demographic information for its predictions takes advantage of data that most pollsters collect anyway and which happen to be good predictors of turnout and over-reporting. Because the PGaD model does make use of this additional information, it not only produces minimal bias, but in our simulations it also produces vote share estimates that are less variant across repeated samples. In other words, PGaD offers reduced bias and increased stability over traditional likely voter models. Furthermore, this method is relatively simple to implement – while we present a random forest approach in the paper, one can also produce the predictions with a basic logit or probit model.[6] Ultimately, our findings suggest that future election predictions can be made more accurate by taking a more information rich modeling approach to identifying likely voters.

[6] And pollsters need not perform vote validation on their own samples to inform these predictions, as publicly available surveys such as the CCES and the American National Election Study now routine include vote validation for respondents.

References

American Association for Public Opinion Research. 2017. "An Evaluation of 2016 Election Polls in the U.S." Ad Hoc Committee on 2016 Election Polling.

Ansolabehere, S., & Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, *20*(4), 437-459.

Ansolabehere, S., & Schaffner, B. F. COOPERATIVE CONGRESSIONAL ELECTION STUDY, 2016: COMMON CONTENT [computer file] release 2: August 4, 2017. Cambridge, MA: Harvard University [producer].

Berent, M., Krosnick, J. A., & Lupia, A. (2011). The quality of government records and over-estimation of registration and turnout in surveys: Lessons from the 2008 ANES Panel Study's registration and turnout validation exercises. Working Paper no. nes012554. Ann Arbor, MI, and Palo Alto, CA: American National Election Studies.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly*, 65(1), 22-44.

Blais, A. (2006). What affects voter turnout? *Annual Review of Political Science*, *9*, 111-125.

Bomey, N.(2016). How did pollsters get Trump, Clinton election so wrong? *USA Today*, November 9. `https://www.usatoday.com/story/news/politics/elections/2016/2016/11/09/pollsters-donald-trump-hillary-clinton-2016-presidential-election/93523012/`

Chalabi, M. (2016). Why were the election polls so wrong? How Donald Trump defied prediction. *The Guardian*. November 9. `https://www.theguardian.com/us-news/2016/nov/09/donald-trump-exit-polls-data-us-election`

Cohn, N. (2016). We gave four good pollsters the same raw data. They had four different results. *NewYorkTimes.com, The Upshot.*

Crespi, I. (1988). *Pre-election polling: Sources of accuracy and error.* Russell Sage Foundation.

Freedman, P., & Goldstein, K. (1996). Building a probable electorate from preelection polls: A two-stage approach. *Public Opinion Quarterly*, *60*(4), 574-587.

Gelman, A., & King, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, *23*(4), 409-451.

Gross, J. (2016). How to better communicate election forecasts – in one simple chart. *Monkey Cage, WashingtonPost.com*, November 29

Keeter, S., Igielnik, R., & Weisel, R. (2016). Can Likely Voter Models Be Improved? Evidence from the 2014 US House Elections. Pew Research Center Report. `www.pewresearch.org/2016/01/07/can-likely-voter-models-be-improved`.

Kiley, J., & Dimock, M. (2009). Understanding likely voters (Methodological Note). Pew Research Center.

Leighley, J. E., & Nagler, J. (2013). *Who votes now?: Demographics, issues, inequality, and turnout in the United States.* Princeton University Press.

Murray, G. R., Riley, C., & Scime, A. (2009). Pre-election polling: Identifying likely voters using iterative expert data mining. *Public Opinion Quarterly*, *73*(1), 159-171.

Newport, F. (2000). How do you define ”likely voters”? Retrieved September 11, 2017, `http://www.gallup.com/poll/4636/how-define-likely-voters.aspx`

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. d., Durand, C., Franklin, C., et al. (2016). An evaluation of the 2016 election polls in the U.S. American Association of Public Opinion Research.

Rogers, T., & Aida, M. (2014). Vote self-prediction hardly predicts who will vote, and is (misleadingly) unbiased. *American Politics Research*, *42*(3), 503-528.

Rolfe, M. (2012). *Voter turnout: A social theory of political participation.* Cambridge University Press.

Silver, N. (2017). The media has a probability problem: The media's demand for certainty — and its lack of statistical rigor — is a bad match for our complex world. *Fivethirtyeight.Com*, September 21.

Tamman M. and Faulconbridge, G. (2016). How the polls, including ours, missed Trump's victory. `https://www.reuters.com/article/us-usa-election-polls/how-the-polls-including-ours-missed-trumps-victory-idUSKBN134306`

Verba, S., & Nie, N. H. (1972). *Participation in America: Political Democracy and Social Equality.* Harper & Row.

Wolfinger, R. E., & Rosenstone, S. J. (1980). *Who Votes?* Yale University Press.