Non-Representative Surveys: Modes, Dynamics, Party, and Likely Voter Space

Tobias Konitzer[a], Sam Corbett-Davies[a], David Rothschild[b]

[a]*Stanford University, Stanford, CA, USA*
[b]*Microsoft Research, New York, NY, USA*

Abstract

State-level forecasts of the popular vote in the 2016 presidential election experienced a large, correlated error in a number of ultimately pivotal states. These forecasts were created by interpreting and aggregating traditional, representative surveys by major news websites. At the same time, we collected responses in a very different kind of non-representative survey: weekly random draws from a non-representative mobile-only panel. Building on work by Wang et al. (2015), we used this mobile-phone-only data in combination with high end analytics to develop the first 51-state projection based on mobile-data only we are aware of. Specifically, we ran our repeated cross-section through dynamic MRP+ (Modeling and Post-stratification). On the survey side, we model the probability that any random respondent, with any combination of specified demographics, would vote for Hillary Clinton, Donald Trump, or other. On the projection side, we use a triangulation of a full updated voter file, Census data, and other historical snapshots of the electorate, to project those probabilities onto an estimated voter space for 2016. Our approach has significant advantages compared to previous MRP-based predictions of elections. We (1) develop a fully dynamic model to disentangle changes in sample composition over time from true swings and compensate for smaller sample sizes in some sub-groups, (2) expand on measures to correct for partisan response bias, (3) estimate the likely voter population directly, instead of relying on naive estimators of previous turnout through exit polls, general population through census data, or demonstrating ex-post predictions based on actual turnout. We show that (a) our forecasting model on a single poll is at least as accurate (if not more) as poll aggregations of public traditional polls, and (b) magnitudes cheaper to conduct. In closing, we discuss ways to quantify uncertainty in our methodology, and elaborate on the future of forecasts based on non-representative polls.

Keywords:

Non-representative polling, multilevel regression and post-stratification, election forecasting

Email addresses: tobiask@stanford.edu (Tobias Konitzer), scrobett@Stanford.edu (Sam Corbett-Davies), davidmr@microsoft.com (David Rothschild)

Introduction

Public opinion polls have been central to election forecasting since the early 20th century. As early as 1916, the magazine *Literary Digest* undertook large-scale, opt-in straw polls and reported the raw percentages as a prediction of the winner of the presidential election. In 1936 the magazine predicted a Republican landslide based on two million mail-in surveys returned to the magazine from its readers and other available lists of voters (mostly automobile and telephone owners, thus skewing Republican), but Democratic President Roosevelt got reelected comfortably instead. After that election, non-representative polls quickly gave way to considerably smaller but representative polls. Specifically, the media industry settled on probability-based polling as the only acceptable form of polling by the 1956 election, especially after quota-based polling was largely seen as responsible for the Dewey v. Truman polling disaster of 1948 (Konitzer and Rothschild, 2016; Wang et al., 2015).

These traditional representative surveys have had respectable accuracy in predicting the results of the popular vote, especially if multiple polls are aggregated. For example, Erikson and Wlezien (2012a) estimate that the average of all polls during the final week of the campaign between 1936 and 2008 deviate from the true outcome with a Root-Mean-Squared Error (RMSE) of only 2.72 percentage points (see also Erikson and Wlezien, 2012b). Likewise, traditional representative surveys did not do too badly in forecasts of the 2016 presidential election. Polling aggregators, web sites that aggregated the publicly available surveys and range from simple averages of the current snapshot (e.g., RealClearPolitics) to complicated algorithms of the actual vote that may include some fundamental data along with the survey data (e.g., FiveThirtyEight), all showed that Democratic candidate Hillary Clinton would win the national popular vote by between 3 and 4 percentage points – a fairly accurate prediction of her 2.1 percentage point margin over Republican candidate Donald Trump at the ballot box.

This level of accuracy mostly held at the state-level. The average error among 15 pre-

determined "toss-up" states was just 2.9 percentage points at RealClearPolitics – a simple averaging of respected polls. But aggregations of representative polls come with a major weakness. They rely on publicly available data, which means they rely on top-line estimates of polling data, which are a combination of several hundred responses and an opaque aggregation method. With no access to the underlying individual-level data of the polls, it is hard for aggregators to understand error and its correlations between polls and elections. Thus, it is unsurprising that poll aggregators produced sizable misses in state-level predictions that a) occurred in contiguous states, and b) were correlated in their errors. RealClearPolitics, which was the most favorable aggregator to Donald Trump was off by 6.6 points (IA) 6.4 points (OH), 3.1 points (PA), 7.5 points (WI), and 3.7 points (MI) respectively *in the same direction*, i.e. favoring Trump.[1] Among other aggregators, these errors were even more pronounced.

In this paper, we build on work by Wang et al. (2015) and leverage responses from a very different kind of non-representative survey: weekly random draws from a non-representative mobile-only panel, in combination with high end analytics, to show that single, non-representative polls can reach the level of accuracy of aggregations of representative polls, and can clearly add meaningful additional information. We develop the first 51-state projection based on mobile-data only we are aware of. Our method includes running our repeated cross-section through dynamic MRP+ (the next generation of Modeling and Post-stratification). Specifically, we model the probability that any random respondent, with any combination of specified demographics, would vote for Hillary Clinton, Donald Trump, or other with our survey data. We then project these probabilities onto our best estimate of the likely voter population – derived from a triangulation of voter file data, Census data and other historical snapshots of the electorate. Compared to Wang et al. (2015), our approach features a number

[1] `http://www.realclearpolitics.com/articles/2016/11/12/it_wasnt_the_polls_that_missed_it_was_the_pundits_132333.html`

of advantages: (1) Instead of relying on exit polls of previous elections, or limiting ourselves to "ex-post predictions" to be able to leverage the exit poll of 2016 (Gelman et al., 2016), we estimate the likely voter space directly; (2) Instead of moving averages, our fully dynamic MRP+ approach is able to disentangle changes in sample composition over time from true changes over time – as far as we know also a novelty; (3) we improve upon ways to correct for partisan response bias, found to contaminate non-probability polls (Gelman et al., 2016), by triangulating individual-level data on ideological leanings from the voter file with aggregate polls of party identification; and (4) we develop a realistic framework of uncertainty of our predictions. In a difficult-to-predict election, our non-representative survey data provided more insight than the traditional representative survey data, indicating, at any point, that either one or two of the pivotal states of Michigan, Pennsylvania, and Wisconsin would go to Republican candidate Donald Trump.

This paper shows that methods can evolve to provide more generalizable solutions for non-representative surveys. Non-representative surveys, with modeling and post-stratifying, allow for timely, flexible, and cheap views of the voting population. The framework we introduce in this paper will allow researchers to have more accuracy and time granularity in assessing public opinion of political events, at a lower cost, leading to new answers to new questions. Polling results were released during the election cycle in order to avoid any look-ahead bias; we made some small methodological changes ex-post which take advantage of thing we learned, but they do not substantially shift the accuracy of the forecast and are noted in the paper.

Mobile Mode, Dynamic Models, and Likely Voter Space

In 2012, Wang et al. (2015) used the Xbox gaming system to demonstrate that opt-in non-representative polling could provide accurate predictions for state-level election outcomes, while being timely, flexible, and cheap. For example, large amounts of data can be collected

much faster than via means of representative polling, and graphical interfaces can increase the flexible of questions served to respondents. As Wang et al. (2015) showed, such data can be leveraged to create an accurate forecast of popular votes in 51 states (50 states plus DC).

In this paper, we replicate this 51-state prediction, leveraging data collected via a different mode – mobile only. The mobile mode comes with a host of advantages. While the set of respondents who can be reached by mobile mode surveys is smaller and less representative of the population, the rich passive data, for example information on installed applications, or precise geolocation, can offer advantages in some research contexts. While survey modes will continue to evolve, we are confident that mobile surveys, with their advantages in accessibility and recent growth in reach, will become a fundamental polling mode, lending the approach presented here added relevancy (Konitzer et al., 2017a).

In addition, we improve upon several limits in Wang et al. (2015). First, Wang et al. (2015) used a sliding window, or moving average, combining the responses from multiple days to stabilize estimates. Second, the authors use party identification as a stable variable, assuming a distribution of partisans that does not change in the course of 4 years. Last, instead of estimating the current turnout population, they relied on the 2008 exit polls. Exit polls are a) polls themselves and subject to considerable potential bias,[2] and b) the breakdown of the turnout population can change considerably over the years (Konitzer et al., 2017b).

This paper builds on Wang et al. (2015) by advancing all three of its key limits. On the survey side, we derive a model that dynamically incorporates each new observation. The intuition behind our dynamic model is simple: Well-identified predictors we use to estimate sub-demographic estimates of vote choice from the survey data are allowed to evolve, with necessary limitation to reduce noise. In contrast, we let less well-identified variables, i.e.

[2]https://www.nytimes.com/2016/06/10/upshot/there-are-more-white-voters-than-people-think-thats-good-news-for-trump.html?˙r=0

marginal demographic "buckets" for which we have a limited sample size in any given wave, be defined by the bulk of historical data. In practice, this model allows us to systematically parse out compositional changes in the sample from true swings. Intuitively, if our estimates for a certain demographic bucket, say White females with a BA degree who are married and identify as Democrats, vary widely from wave to wave, our model allows us to understand how much of this change is due to sample composition, and how much is due to attitudinal swings. We describe the model in full below. Second, we improve precision in our ability of correcting for partisan response bias (Gelman et al., 2016). Instead of relying on a single measure of self-reported party id, we make use of feeling thermometers to reduce measurement error. Specifically, we remove respondents who identify as strong partisans but indicate a score of 10 or less toward the in party. This kind of measurement error can be common in digitally administered polls such as ours or the one analyzed in Wang et al. (2015).

In contrast to Wang et al. (2015), we also derive a predicted turnout, or voter space. We start with the most recent 5-year-estimate from the 2015 Census American Community Survey.[3] This provides us a very accurate view of the demographics of any district or state, but does not include anything about their ideology or likeliness to vote. So, we impute turnout probability and ideology score (0=conservative to 100=liberal) from the full voter file by demographic. Ideology is predicted by projecting 10,000s of survey responses on policy preferences onto each demographic combination in the file, allowing us ultimately to leverage individual-level ideology scores. Turnout is mostly taken from previous turnout history, and modelled based on survey data for first-time voters. Specifically, we weight the cell-counts for each sub-demographic by their predicted turnout probability, and impute a party identification score to each sub-demographic, using our continuous individual-level ideology score. We bucket these continuous scores into Democrat, Republican and Independent

[3]https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2015/5-year.html

identification such that the marginals in our likely voter space match the marginals of the national party identification distribution taken from Huffington Post Pollster.

We could leverage the full voter file exclusively to derive the voter space; in practice, this is not an advisable approach. Voter files are designed with individual-level accuracy in mind. Some of the demographic variables that partly rely on imputation or predictions are prone to aggregation errors. For example, voter files might overstate the proportion of White voters because individual-level predictive accuracy of race can be maximized by over-predicting Whites. However, if voter files list the most *likely* racial affiliation of a voter instead of the full probabilistic breakdown, correlations between prediction errors increase exponentially when aggregated. In consequence, we prefer to use reliable data on the demographic breakdown of the US population as a baseline.

This explicit approach also allows us to address the margin of error (MoE) of our estimates. As Shirani-Mehr et al. (2017) recently discussed, the margin of error in traditional polls vastly understated uncertainty in recent elections, because the standard MoE is driven by sample size and does not properly account for sample quality. In our approach, we note that overall error can be decomposed into three sources: Model error – capturing (a lower baseline of) sampling error, coverage error and response error – measurement error, and projection space error, i.e. errors in estimating likely voters. While our methodology captures model error naturally, we cannot differentiate well between measurement error and projection space error. Given that we have taken precautions against measurement error, however, we can conduct ex post calibration to get at a (first) estimate of projection space error.

Data

Pollfish is a mobile-based polling platform that allows app developers to monetize their apps by including pop-up surveys in lieu of traditional ads (Figure 1). Pollfish can thus potentially survey every user of the hundreds of apps they partner with. Third-party appli-

cations display advertisements to users, and some of those contain a Pollfish poll request. When a user clicks the request, the Pollfish application launches (see Figure 1). Akin to other panels,[4] Pollfish collects demographic information – gender, age, education, race, income, and other household-level information – via a demographic screen. In the US, Pollfish has 10 Million monthly active registered users. Globally, the number of active users is greater than 300 million.[5] Of course, only a tiny fraction respond to any survey, and post-survey analytics are required to produce representative estimates due to the opt-in character of this panel.

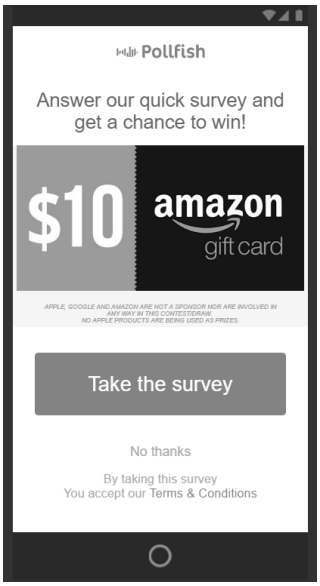

Figure 1: Pollfish Poll Request

Beginning January 2016, we surveyed Americans once a month on their preferred presidential candidate (Appendix 1). Before Donald Trump and Hillary Clinton had clinched their parties' nominations, we asked whether voters would prefer the Republican candidate or the Democratic candidate. As the head-to-head campaign began in earnest after the conventions, we began to poll more frequently, peaking at 3 polls in the week before the

[4]See for example `https://today.yougov.com/`.
[5]`http://www.pollfish.com`

election. In total, we received 26,000 responses over 11 months that we collapse to 21 waves. Along with their vote intention, we received each respondent's age, gender, race, education level, state of residence and marital status through the demographic screen, and partisan identification from our poll.

Our poll resembles a repeated cross-section, emphasizing the need for dynamic modeling. Post-stratification can address some survey error, but – conditional on the observed demographics – the respondents on a given day may still be indicative of partisan imbalance. For example, it is conceivable that strong Clinton supporters were more likely to respond to political polls after a strong debate performance by their preferred candidate (e.g. Gelman et al., 2016),[6] and that this difference in sample composition is not entirely captured by party identification. Our model smooths out these day-to-day variations in our sample frame to provide an estimate of meaningful swings parsed from sample compositional differences over time.

To derive our likely voter space, we rely on the American Community Survey (ACS), a large national sample that gives much of the same information on demographic breakdowns as the long-form Census. Specifically, we leverage all US citizens from the ACS 5-year-estimates 2011-2015 (N=2,469,680 for voting-age citizens). To get measures of turnout and party identification, which are not available in the ACS, we rely on the complete 2016 VoterBase voter file, compiled by TargetSmart. These files encompass over 146 million registered voters in the United States, including address, party registration, gender, age and other demographic variables as well as machine-learning-based predictions for ideology and probability of turnout in the 2016 presidential election. This data set allows us to impute party and turnout propensity by demographic subgroup, as described above. Finally, we rely on Huffington Post Pollster estimates for the distribution of party identification to calibrate

[6]https://www.washingtonpost.com/politics/as-clinton-builds-on-a-strong-debate-trump-lobs-attacks-and-complaints/2016/09/27/6bb4cd2e-84cc-11e6-92c2-14b64f3d453f˙story.html?utm˙term=.c7a1a264ea4c

our party scores, as likewise described in the previous section.[7]

Estimation Strategy

Modeling and Post-stratification

For any single cross-sectional poll (in our models denoted as *wave*, and indexed with t), our strategy for estimating vote intention follows Wang et al. (2015). Using the rich set of demographic variables, we split the respondents into M demographic cells and estimate the two-party vote share in each cell.

The set of cells is the Cartesian product of all age, gender, race, education level, marital status, partisan identification, and geographic division categories[8], over one thousand in total. With an abundant number of cells, estimating a cell's two-party vote share by simply averaging responses of the survey respondents in that cell produces highly unstable estimates (and is impossible in the majority of cells which are empty in a given wave). In general, approaches such as simple means, raking or post-stratification weighting suffer from a Bias-Variance trade-off. Including more variables to address survey error inflates the number of cells exponentially. In the most extreme cases, cells become empty, rendering the post-stratification estimator, or mean estimator, undefined. But even in more realistic cases, the influence of single observations increases sharply, introducing variability that increases the total survey error. For example, a single observation in a USC/LATimes election poll – a 19 year-old black man from Illinois – who uncharacteristically intended to vote Republican was weighted 300 times more than the least weighted respondent, changing the top-line estimate

[7] http://elections.huffingtonpost.com/pollster/party-identification-voters

[8] The possible categories are as follows. Age is divided into 5 categories: 18-24, 25-34, 45-54, and 55+. Race is self-reported as white, black, Hispanic, and other. Education level divides those with at least a bachelors degree from those without. Partisan identification has 3 levels: Democrat, Republican, or Independent. Geographic division uses the US Census Bureau's classification of states into 9 divisions.

of that poll by 2 percentage points.[9] On the other hand, not including variables that govern survey error and at the same time are related to the outcome introduces bias.

We overcome this trade-off by modeling the survey responses in terms of demographics, "borrowing" responses from demographically similar cells.

Formally, we estimate Clinton's share of the two-party vote in demographic cell i and wave t (denoted $\hat{y}_i^t$) using the following logistic regression model:

$$\begin{aligned}\hat{y}_i^t &= P(Y_i^t = \text{Clinton} \mid Y_i^t \in \{\text{Clinton, Trump}\}) \\ &= \text{logit}^{-1}\Big(\alpha^t_{\text{gender}}[\text{gender}_i] + \alpha^t_{\text{age}}[\text{age}_i] + \alpha^t_{\text{race}\times\text{edu}}[(\text{race}\times\text{edu})_i] + \alpha^t_{\text{edu}}[\text{married}_i] + \\ &\qquad \alpha_{\text{division}}[\text{division}_i] + \alpha_{2012} \times (\text{2012 Obama vote})_i + \alpha_{\text{order}} \times (\text{survey order})_t\Big) \qquad (1)\end{aligned}$$

where $\alpha^t_{\text{demo}}[k]$ is the random effect parameter corresponding to the k'th level of demographic *demo*, and demo_i is the value of *demo* in demographic cell i. The parameters are estimated in a hierarchical Bayesian framework, which we elaborate on in the following subsections.

The model specification in Eq. 1 closely follows that used by Wang et al. (2015), with the addition of the survey order variable. This was introduced after conducting the first month of surveys, when we noticed that the level of support for Trump among self-identified "Strong Democrats" was surprisingly high (although not high in an absolute sense). Further investigation revealed that this was due to measurement error: a small proportion of respondents were selecting the top answer to most questions. Since we didn't randomize the vote preference options (in order to keep the major party candidates at the top of the list), Trump—as the top option—had his support inflated. Rather than remove these surveys, we added a survey order dummy that varied by survey wave to capture the extent of the measurement error. During model fitting this dummy took a value of 1 if Trump was listed

[9] See https://www.nytimes.com/2016/10/13/upshot/how-one-19-year-old-illinois-man-is-distorting-national-polling-averages.html.

first, and 0 if Clinton was. When computing our estimates of the cell means $\hat{y}_i^t$ the dummy was set to 0.5, averaging out the survey order effects.

Given $\hat{y}_i^t$, we can compute an estimate of Clinton's share of the two party vote among a group of interest G by weighting each $\hat{y}_i^t$ by N_i the number of voters in that demographic cell derived from our estimated likely voter space:

$$\hat{Y}_G^t = \frac{\sum_{i \in G} N_i \hat{y}_i^t}{\sum_{i \in G} N_i}. \tag{2}$$

Common groups of interest are all voters ($G = \{1, ..., M\}$), and voters in state s ($G = \{i \mid \text{state}_i = s\}$), but Clinton's share of the two party vote among any demographic subgroup can be estimated in this way.

Dynamic Modeling

It is rare for polling firms to conduct only a single poll ahead of a major election. Public opinion is usually tracked many months in advance to determine the "state of the horse race" and identify important events during the campaign. However, most analytic strategies analyze each poll in isolation, ignoring this temporal structure. Such tracking polls can exhibit large swings – for example, the Gallup tracking poll in 2012 estimated a highly unlikely increase in Obama vote intention from 44% to 50% in the course of one week alone[10] – and there is concern that this is due to partisan non-response (Gelman et al., 2016). This occurs when partisans, encouraged (or discouraged) by recent news, become more (or less) willing to participate in political polls without changing their vote preference or propensity to vote. Typical analytic strategies will measure this differential non-response as a change in preferences. Gelman et al. (2016) attempted to address this by including attitudinal variables in the post-stratification space – including partisan and ideological affiliations – and found

[10] http://www.gallup.com/poll/150743/obama-romney.aspx

that the swings in the 2012 election were mostly artifacts of partisan non-response.

Nonetheless, swings still occur *within partisan buckets*. How can a pollster determine whether these reflect real changes in opinion, or merely differential non-response *within* partisan categories? For example, Republican-leaning independents may have been less likely to respond to polls after the Democratic National Convention. In this section we present a dynamic modeling and post-stratification method that attempts to parse out changes in opinion from changes in sample composition.

The key assumption underlying our method is that, while shocks to public opinion can occur, the average preferences of any group is relatively stable week-to-week. Panel surveys have shown that very few voters report changing their candidate preference during a campaign (Hillygus and Jackman, 2003), and even fewer switch between major party candidates (Gelman et al., 2016). Thus, the random effect parameters $\alpha^t_{\text{demo}}[k]$ should remain stable between any pair of waves t and $t+1$. We encode this assumption by constraining the parameters to evolve according to an auto-regressive AR(1) process:

$$\alpha^{t+1}_{\text{demo}}[k] = \mu_{\text{demo}}[k] + \phi_{\text{demo}}[k](\alpha^t_{\text{demo}}[k] - \mu_{\text{demo}}[k]) + \epsilon^{t+1}_{\text{demo}}[k]. \tag{3}$$

At each wave, the parameter estimates from the previous wave are shrunk towards the mean $\mu_{\text{demo}}[k]$ by a factor $\phi_{\text{demo}}[k]$, before mean-zero noise $\epsilon^{t+1}_{\text{demo}}[k]$ is added.

We complete the Bayesian specification of the model with the following priors:

$$\begin{aligned}
\alpha^1_{\text{demo}}[k] &\sim \text{student-t}(\mu_{\text{demo}}, \frac{\sigma_{\text{demo}}}{\sqrt{1-\phi^2_{\text{demo}}}}; \nu) \\
\sigma_{\text{demo}} &\sim N_+(0, 0.1) \\
\phi_{\text{demo}} &\sim \text{Beta}(10, 1) \\
\mu_{\text{demo}} &\sim N(0, 1) \\
\epsilon^t_{\text{demo}} &\sim \text{student-t}(0, \sigma_{\text{demo}}; \nu).
\end{aligned}$$

Drawing ϕ_{demo} from a beta distribution ensures that $0 < \phi_{\text{demo}} < 1$, meaning that the AR(1) process determining the evolution of the parameters is stationary, with mean μ_{demo} and variance $\frac{\sigma^2_{\text{demo}}}{1-\phi^2_{\text{demo}}}$. The initial parameter values $\alpha^1_{\text{demo}}[k]$ are also drawn from the stationary distribution.

Stationarity is important because we have no reason, *a priori*, to believe that the parameters will exhibit a long-term temporal trend. Of course, the *posterior* distribution could still exhibit such a trend (e.g. whites could become increasingly supportive of Trump), if justified by the polling data.

The hyperparameters σ_{demo}, $\mu_{\text{demo}}[k]$, and $\phi_{\text{demo}}[k]$ are estimated using the data from all waves. Intuitively, $\phi_{\text{demo}}[k]$ measures the autocorrelation of the preferences of a given demographic, while σ_{demo} captures how much these preferences vary around the mean. This allows us to identify which covariates are stable predictors of vote intention and which are susceptible to the ebbs and flows of the campaign. For example, the residual explanatory power of a voter's state tends to be very stable, with white Democrats in Kentucky expressing consistently more conservative preferences than white Democrats in Maine. Conversely, the effect of partisan identification is more variable, as Independents and disaffected partisans consider alternative options (such as 3rd party candidates) at different points in the race.

The choice of the ν in the student-t distribution is important to determine whether changes in public opinion occur smoothly or in response to shocks. A small ν value puts more mass in the tails of the distribution, favoring larger shocks over smooth movements. We choose $\nu = 4$.

Bayesian inference

We estimate the specified model in a Bayesian framework. Specifically, we fit the model using Stan (Sampling Through Adaptive Neighborhoods) with Hamiltonian Monte Carlo (HMC) implemented via the No-U-Turn (NUTS) algorithm in C++. The full code is included in Appendix 2. We sample 1,000 draws from 4 chains of the posterior, allowing for a warm-up, or burn-in of 1,000 iterations. Each sample is a plausible realization of the parameters, and implies a certain vote share for Clinton in the group of interest. We use the median of the vote share distribution as our point estimate, and the 95% credible interval (between the 2.5 and 97.5 percentiles) to describe model uncertainty.

Results

We begin by displaying our top-line estimates going into election day. On November 05, three days before the election and on the day of our last wave, we have Clinton leading over Trump – 50.56% vs.49.44% (two-party vote share). Ultimately, Clinton took 51.11% of the two-party vote, and Trump took 48.89%. At the national level, our final estimates were hence off by less than 0.5 percentage points, and en par with many other respectable polling aggregators (Kennedy et al., 2017).

We also document some interesting movements over time. While we discuss our dynamics in more details below, we note here that while in our model, Clinton was in the lead for most of the campaign cycle, and certainly for the last two months, we had her trailing Donald Trump briefly in January and between August and September.

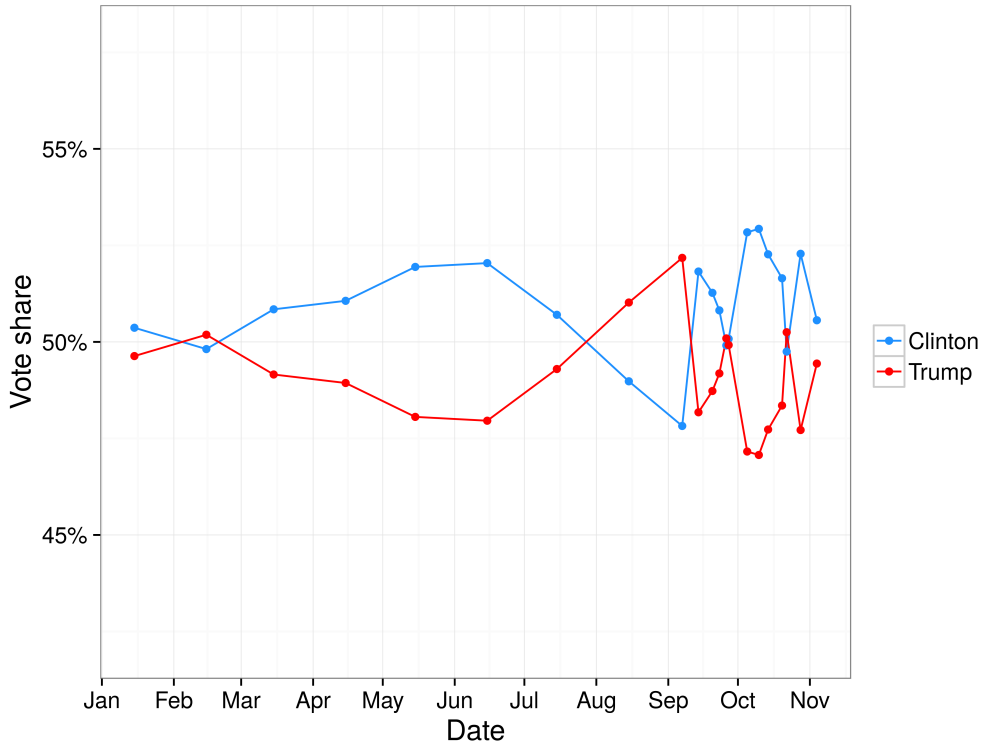

Figure 2: Predicted Two-Party Vote Share in the 2016 Presidential Election

Next, we validate our subgroup estimates against exit polls. We note that exit polls do not represent the ground truth, but are a poll in and of themselves.[11] However, the general convergence of our predictions for sub-demographics of interest – specifically race x gender interactions, education, race, gender, party and married – with exit polls estimates are quite strong. Besides our predictions for black females and other race, all estimates are within 5 percentage points from exit poll estimates. Note that the two categories where we do see divergence are those that are the *least* stable in modern exit polls.[12]

[11] https://www.nytimes.com/2016/06/10/upshot/there-are-more-white-voters-than-people-think-thats-good-news-for-trump.html

[12] https://www.nytimes.com/2016/06/10/upshot/there-are-more-white-voters-than-people-think-thats-good-news-for-trump.html

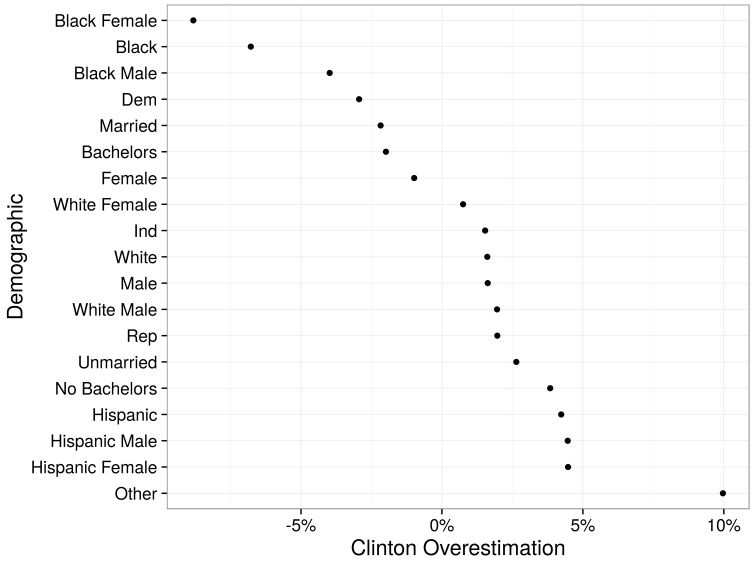

Figure 3: Predicted Two-Party Vote Share vs. Vote Share Estimated by Exit Polls for Sub-demographics of Interest

Next, we present our final state-by-state predictions. We code states in which we have either candidate up by more than 10 percentage points (two-party vote share) as "strong", states in which we have either candidate up by more than 2 percentage points as "lean", and all other states as toss-up.[13] Our results indicate Pennsylvania, Florida and North Carolina to be tilting Republican, with Ohio, Virginia and Michigan as true toss-ups. This is strikingly different from polling aggregators. The model of *The New York Times* for example, in their polling aggregation, had Clinton ahead in Pennsylvania, Florida and North Carolina by 4.1 percentage points, 2.2 percentage points and 2.3 percentage points respectively. In total, the only states we missed (in a binary sense) were New Hampshire Maine, which Clinton carried by 0.37 and 2.96 percentage points respectively.

[13]See `https://www.pollfish.com/blog/2016/11/11/pollfish-presidential-election-2016/` for a true prediction of state-by-state results. The final model used in this paper is very similar to the one leading to our true predictions.

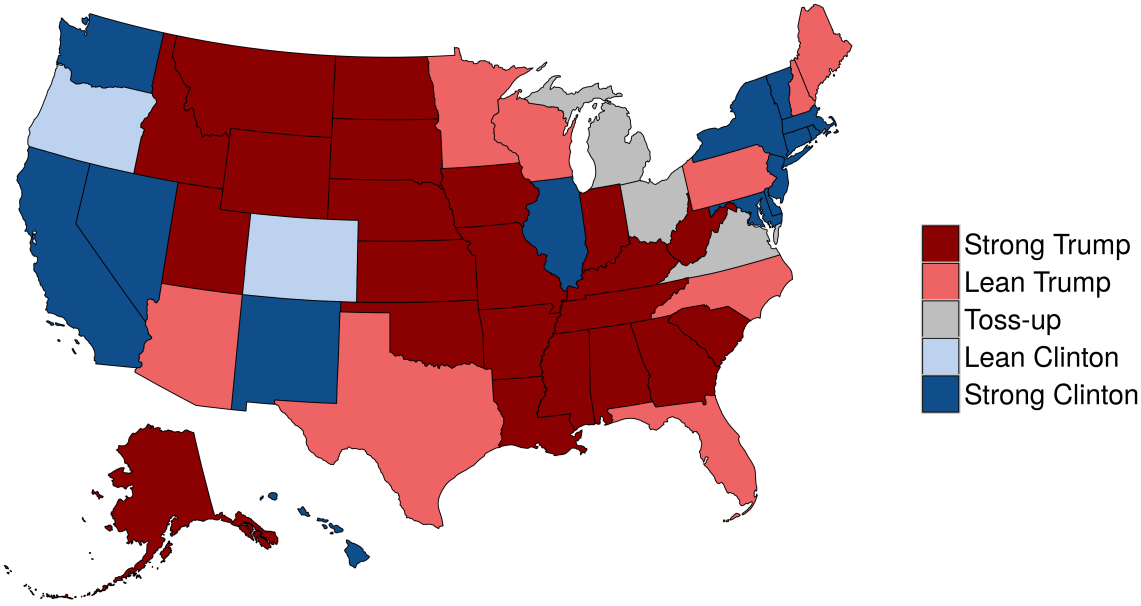

Figure 4: Predicted state-by-state outcome of the 2016 presidential election, as of 2016-11-05. Toss-up states are where the candidates are within 2 percentage points of each other; in leaning states the margin is less than 10 points.

When we try to quantify our state-by-state errors more precisely, we find our predictions based on a single poll do not do significantly worse than the predictions from poll aggregators. Below (Figure 5), we compare our state-by-state estimates against the actual outcome. Compared to poll aggregator Huffington Post Pollster, our Root Mean Squared Error (RMSE) is only slightly higher – 4.24 percentage points vs. 3.62 percentage points (for 50 states excluding DC).

When we focus on the 15 closest states, our predictive accuracy is even higher. Our RMSE is 2.89 percentage points, compared to 2.57 percentage points of Huffington Post Pollster. Overall, besides binary accuracy our predictions also have low error in the precise percentage value.

Figure 5: Predicted State-by-State Outcome of Presidential Election 2016 vs. actual State-by-State Outcome of Presidential Election 2016 for all states except DC (left panel); and for 15 closest States (right panel)

Not only are our state-by-state estimations fairly accurate, they also add meaningful signal to the poll aggregations. The left panel of Figure 6 displays the correlation between state-by-state errors of our predictions and the state-by-state errors of Huffington Post Pollster, and the right panel compares the distribution of errors across our approach and Huffington Post Pollster. At the very least, including data sources such as ours has significant potential to increase the quality of aggregators, as we discuss more below.

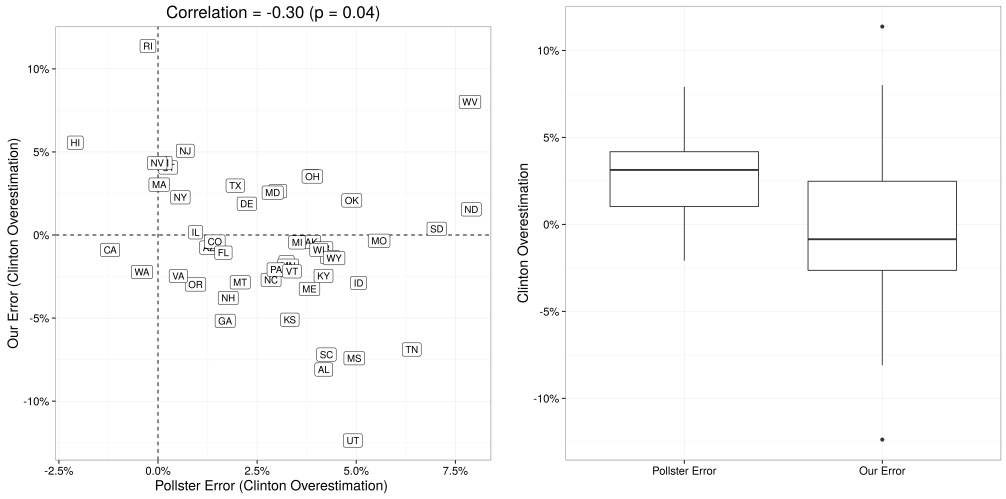

Figure 6: Correlation of Errors in our State-by-State Prediction and Huffington Post Pollster's Estimates (left panel); Distribution of Errors in State-by-State Predictions in our Approach and Huffington Post Pollster's Estimates (right panel)

Next, we look at our dynamics. We display our estimates of two-party vote share, again compared to Huffington Post Pollster's. First, we note that our estimates are almost as stable as those of aggregated high quality polls. Second, some of the observed movements correlate with anecdotal evidence from the campaign trail. For example, the bounce our model registers for Hillary Clinton after the first presidential debate (first dashed line in Figure 7) corresponds with journalistic takes that generally concluded Clinton outperformed Trump.[14] While effects of the second and third debate (second and third dahsed line) appear small, the Comey letter (fourth dashed line), indicating a reopening of the investigation into Clinton's use of a private email server,[15] caused a significant drop in Clinton support in our model, although there appears to be some lag.

[14]http://www.cnn.com/2016/09/27/politics/presidential-debate-hillary-clinton-donald-trump-highlights/index.html

[15]https://www.washingtonpost.com/politics/fbi-to-conduct-new-investigation-of-emails-from-clintons-private-server/2016/10/28/0b1e9468-9d31-11e6-9980-50913d68eacb˙story.html?utm˙term=.0720106a6455

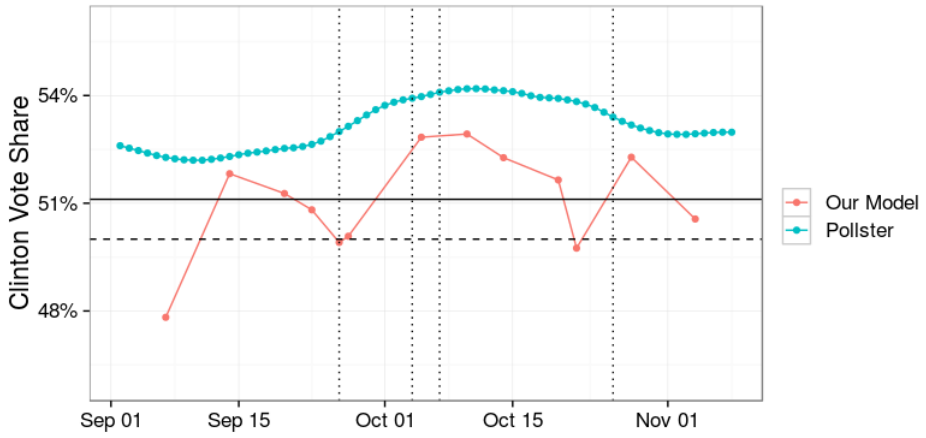

Figure 7: Dynamic two-party vote share in our approach and in Huffington Post Pollster

On another front, our dynamics likewise support anecdotal journalistic evidence from the campaign trail – the drop in Clinton support by uneducated Whites.[16] The decision to include a race x education interaction was made ex post, i.e. after the election (and is the only difference between our true prediction models available at `https://www.pollfish.com/blog/2016/11/11/pollfish-presidential-election-2016/` including time stamp, and our ex post models). The drop in support was somewhat higher for uneducated than for educated whites. This finding casts doubt on the journalistic narrative that uneducated Whites were always out of reach for Clinton.[17] Instead, our model suggests that the drop in support began in earnest in July.

[16] Unfortunately, we are not able to break down education further

[17] https://www.nytimes.com/2016/07/26/upshot/the-one-demographic-that-is-hurting-hillary-clinton.html

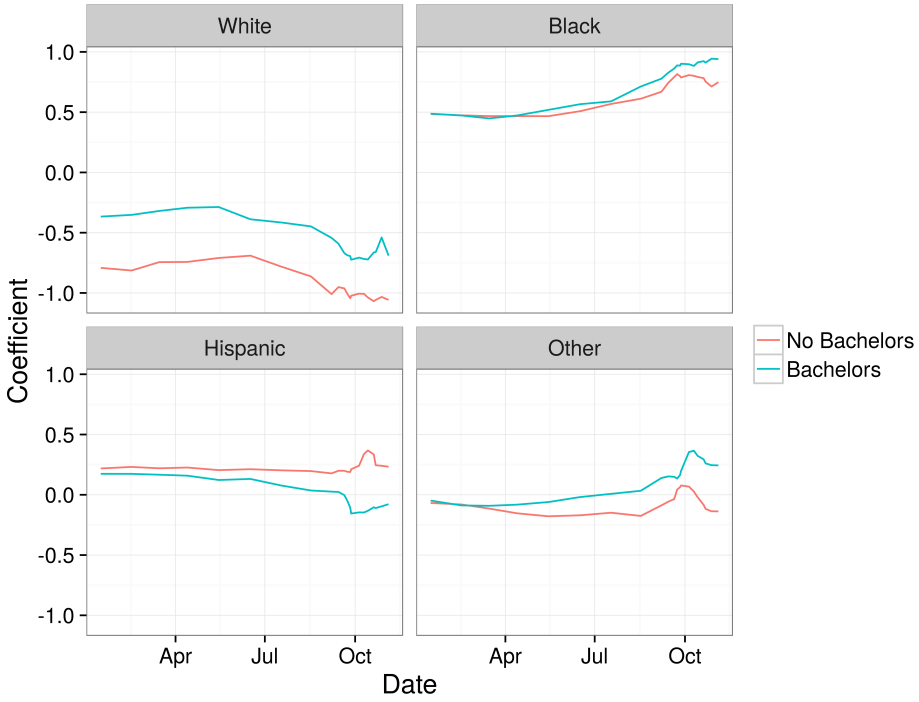

Figure 8: Two-party Vote Share Estimates for Gender x Education Interaction over Time; Positive Coefficients Indicating Support for Clinton, Negative Coefficients Indicating Support for Trump

Last, we look at the uncertainty in our state-level point estimates. As discussed earlier, we note that the overall error in our estimates is a combination of three things: Model error, measurement error and turnout error. The left panel of Figure 9 below displays point estimates including 95% Bayesian credible intervals for state-by-state estimates of the 15 closest states. In the right panel, we have expanded our error bars such that we have 95% coverage, i.e. such that our uncertainty estimates contains the truth for at least 48 of the 51 states. In sum, in our 48^{th} most accurate state the distance from the error bar to the ground truth is 4.8 percentage points, such that if we expanded our errors by that margin, we achieve 95% coverage of the ground truth. Given our discussion above, we have reason to believe that the majority of this error captures uncertainties in turnout.

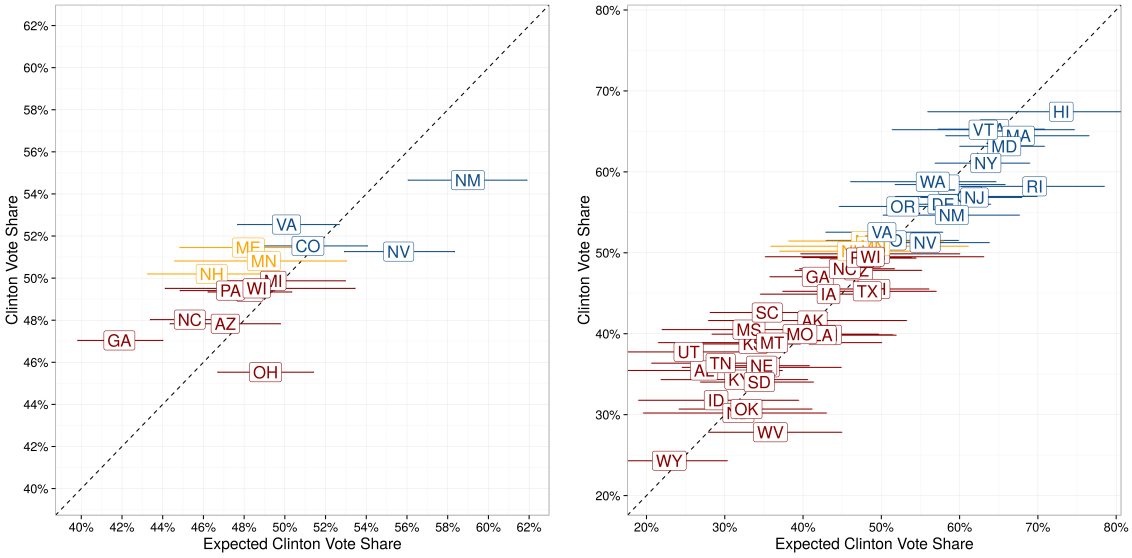

Figure 9: Point-estimates of Projections for 15 Closest States including 95% Bayesian Credible Intervals (left panel); Point-estimates of Projections for all States including Expanded Errors Achieving 95% Coverage of Ground Truth (right panel)

Discussion

This paper shows that non-representative polling data can be used to describe vote intention of the public accurately – even with limited sample size, if we make use of the dynamic data structure. Compared to previous research, leveraging non-representative polling (e.g. Wang et al., 2015), we improve upon core elements. First, we introduce a dynamic model-based post-stratification (MRP+) model that is able to produce stable estimates even under conditions of limited sample size by pooling information across waves and learning the degree of pooling from the data. Second, we account for fluctuations in the likely voter space that occur from election to election (Konitzer et al., 2017b). Third, we advance ways in which to control for partisan non-response bias (Gelman et al., 2016).

By no means do we advocate to replace traditional polls, or even the aggregation of traditional polls, with a single (and in addition non-representative) poll, but our results do show that novel data sources, such as mobile panels leveraged here, can add significant additional signal to the already rich signal of polling aggregators. This is no surprise in and

of itself, given that traditional probability polls might come with mode-specific biases. For example, research firm Civis Analytics points to the possibility of systematic non-response bias among lower educated white rural voters in telephone polls.[18] If that response bias is concentrated among mobile blue color workers, as recently argued by Shor and Swasey (2017), it makes sense that data sources able to collect data more independent of location, such as mobile polls, add signal to traditional polling data.

Another question relates to the uncertainty inherent in our predictions. As Shirani-Mehr et al. (2017) have recently pointed out, the margin of error (MoE) usually reported alongside representative polls is oftentimes too low. Given that precise sampling mechanisms are unknown in all non-representative polls, it does not make sense to derive a similar quantity in our application. Instead, we urge researchers to develop a pendant to the Total Error Perspective dominating the understanding of uncertainty in probability polls (Biemer, 2010). Error in our approach can stem from three sources: a) our model, b) measurement error, and c) voter space error, i.e. error stemming from our estimation of who turned out in 2016 or not. We can address a) by including model error in our estimates, akin to Figure 9. We can now calibrate our state-estimates in that we inflate errors by an added constant such that we achieve desirable coverage. For example, under 95% confidence, we need to inflate our errors by 4.8 percentage points such that 95% of states are predicted correctly. The scalar captures b) and c). While this error is impossible to decompose further into voter space and measurement error, we can reasonably assume that the bulk of it captures voter space error, given our strategy to minimize measurement error common in mobile surveys (Konitzer et al., 2017a). Hence, the upper bound of voter space error is 4.8 percentage points. And while we acknowledge that the quantification of this kind of error a) deserves the attention of future research and b) is election dependent, we offer a first approach of

[18] https://www.wired.com/2016/11/pollsters-missed-bowling-alone-voters-handed-trump-presidency/

decoupling survey errors in the framework of non-probability polls paired with MRP+.

In closing, we note that presenting simple top-lines, as most polls do, makes aggregation fairly difficult. Aggregators have no choice but to simply average over different estimates (i.e. Huffington Post Pollster), or to construct a weighted average (i.e. FiveThirtyEight). The method presented here offers a unified framework under which both, non-probability and probability polls can be combined into single predictions, if individual-level data are made public. The advantages of this approach are obvious. First, by modeling the survey data separately, we can use both data sources to fill sub-demographic buckets more completely. Conditional on having identified the true model governing the outcome, i.e. vote choice, combining data from representative and non-representative polls without taking sample mechanism into consideration is possible. Second, we can leverage survey data without having to rely on the same data to estimate the likely voter space. In total, we are confident that adding more data, regardless of sampling mechanism, to a unified framework will help stabilize estimates and prevent polling failures such as the state-by-state estimates in the Rust Belt in the 2016 presidential campaign.

Biemer, P. P., 2010. Total survey error: Design, implementation, and evaluation. Public Opinion Quarterly 74 (5), 817–848.

Erikson, R. S., Wlezien, C., 2012a. Markets vs. polls as election predictors: An historical assessment. Electoral Studies 31 (3), 532–539.

Erikson, R. S., Wlezien, C., 2012b. The timeline of presidential elections: How campaigns do (and do not) matter. University of Chicago Press.

Gelman, A., Goel, S., Rivers, D., Rothschild, D., 2016. The mythical swing voter. Quarterly Journal of Political Science 1 (1), 103–130.

Hillygus, D. S., Jackman, S., 2003. Voter decision making in election 2000: Campaign effects, partisan activation, and the clinton legacy. American Journal of Political Science 47 (4), 583–596.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, E., Wlezie, C., 2017. An evaluation of 2016 election polls in the u.s.
URL http://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

Konitzer, T., Eckman, S., Rothschild, D., 2017a. Mobile as survey mode. In: Working Paper.

Konitzer, T. B., Rothschild, D. M., 2016. Why political polling is not dead–a plea for non-probability polling, algorithms, and big data. Working Paper.

Konitzer, T. B., Shirani-Mehr, H., Rothschild, D., Goel, S., 2017b. Decoupling sources of electoral change - evidence from recent congressional elections. Working Paper.

Shirani-Mehr, H., Rothschild, D., Goel, S., Gelman, A., 2017. Disentangling bias and variance in election polls. Working Paper.

Shor, D., Swasey, C., 2017. Why nobody saw trump coming: Nonresponse bias among non-college educated whites. CIVIS Analytics, American Association of Public Opinion Research.

Wang, W., Rothschild, D., Goel, S., Gelman, A., 2015. Forecasting elections with non-representative polls. International Journal of Forecasting 31 (3), 980 – 991.

AppendixA. Questionnaire for Pollfish

1) How do you feel about reducing federal spending by replacing Medicare with a voucher program? : Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

2) How do you feel about increasing income taxes for people making over $250,000 per year? : Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

3) How do you feel about a military solution to try to prevent Iran from developing nuclear weapons? : Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

4) The number of immigrants from foreign countries should be... : Increased a lot : Increased moderately : Increased only a little : Left the same as it is now : Decreased only a little : Decreased moderately : Decreased a lot

5) Government regulations for businesses should be... : Increased a lot : Increased moderately : Increased only a little : Left the same as it is now : Decreased only a little : Decreased moderately : Decreased a lot

6) How do you feel about abortions being legal in cases of rape, incest, or threat to the woman's health? : Favor Very Strongly Favor Strongly Favor Weakly Neither Favor nor Oppose Oppose Weakly Oppose Strongly Oppose Very Strongly

7) How do you feel about laws to protect individuals against discrimination based on sexual orientation? : Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

8) How do you feel about government regulation mandating maternity leave? : Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

9) How do you feel about government measures to reduce differences in income levels? :

Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

10) How do you feel about federal laws to make it more difficult for people to buy a gun? : Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

11) Do you agree that human behavior is substantially responsible for Global Warming? : Agree Very Strongly : Agree Strongly : Agree Weakly : Neither Agree nor Disagree : Disagree Weakly : Disagree Strongly : Disagree Very Strongly

12) How do you feel about rolling back free trade agreements? : Favor Very Strongly : Favor Strongly : Favor Weakly : Neither Favor nor Oppose : Oppose Weakly : Oppose Strongly : Oppose Very Strongly

13) How do you feel about the Democratic Party today? : 0 - Hate : 10 : 20 : 30 : 40 : 50 : 60 : 70 : 80 : 90 : 100 - Love

14) How do you feel about the Republican Party today? : 0 - Hate : 10 : 20 : 30 : 40 : 50 : 60 : 70 : 80 : 90 : 100 - Love

15) What is your political party affiliation? : Strong Democrat : Weak Democrat : Lean Democrat/ Independent : Independent : Lean Republican/Independent : Weak Republican : Strong Republican

16) How often do you talk to your social network about the presidential campaign? : Every Day : Several Times a Week : A few times a month : Almost Never

17) Who are you most likely to vote for in the upcoming presidential election? : Definitely Republican candidate : Likely Republican candidate : Likely Democratic candidate : Definitely Democratic candidate : Not voting

AppendixB. MRP model

```
data {
```

```
  real<lower=0> nu;                                          // number of
degree of freedom for evolution-prior

  int<lower=0> N;                                           // sample wave_size
  int<lower=0> n_waves;                                     // number of Waves

  int<lower=1> M_dyn;                                       // Number of poststrat
variables with dynamic effects
  int<lower=1> N_dyn[M_dyn];                                // Number of categories
for each dynamic poststrat variable

  int<lower=1> M_fix;                                       // Number of poststrat
variables with static effects
  int<lower=1> N_fix[M_fix];                                // Number of categories
for each

  int<lower=1> M_cov;

  int<lower=1> X_dyn[N,M_dyn];
  int<lower=1> X_fix[N,M_fix];
  matrix[N, M_cov] X_cov;

  int<lower=0> y[N];                                        // Number of pro-votes
for demographic subgroup
  int<lower=0> total[N];                                    // Number of votes
for demographic subgroup
```

```
  int<lower=0> wave_start[n_waves];
  int<lower=0> wave_end[n_waves];
}

transformed data {
  vector[1] zero;

  zero[1] <- 0;
}

parameters {

  real mu_alpha;

  real<lower=0> scale_sigma;

  vector[N_dyn[1]] beta_0_race_x_education;
  vector[N_dyn[2]] beta_0_gender;
  vector[N_dyn[3]] beta_0_married;
  vector[N_dyn[4]] beta_0_party;

  vector[N_dyn[1]] beta_raw_race_x_education [n_waves-1];
  vector[N_dyn[2]] beta_raw_gender [n_waves-1];
  vector[N_dyn[3]] beta_raw_married [n_waves-1];
  vector[N_dyn[4]] beta_raw_party [n_waves-1];
```

```
  vector[N_dyn[1]] mu_dyn_race_x_education;
  vector[N_dyn[2]] mu_dyn_gender;
  vector[N_dyn[3]] mu_dyn_married;
  vector[N_dyn[4]] mu_dyn_party;

  vector[N_fix[1]] mu_fix_division;

  real <lower=0> sigma_dyn_race_x_education;
  real <lower=0> sigma_dyn_gender;
  real <lower=0> sigma_dyn_married;
  real <lower=0> sigma_dyn_party;

  real <lower=0> sigma_fix_division;

  real <lower=0, upper=1> phi_dyn_race_x_education;
  real <lower=0, upper=1> phi_dyn_gender;
  real <lower=0, upper=1> phi_dyn_married;
  real <lower=0, upper=1> phi_dyn_party;

  real alpha_Obama_voteshare;
  real alpha_meas_error_party_id;
}

transformed parameters {
  real alpha [n_waves];
```

```
  vector[N_dyn[1]] beta_dyn_race_x_education [n_waves];
  vector[N_dyn[2]] beta_dyn_gender [n_waves];
  vector[N_dyn[3]] beta_dyn_married [n_waves];
  vector[N_dyn[4]] beta_dyn_party [n_waves];

  vector[N_fix[1]] beta_fix_division [n_waves];

alpha[1] <- mu_alpha;# + alpha_0 * sigma_alpha_dyn * scale_sigma / sqrt(1 - phi_alpha_dy
* phi_alpha_dyn);

  beta_dyn_race_x_education[1] <- mu_dyn_race_x_education + beta_0_race_x_education
* sigma_dyn_race_x_education * scale_sigma / sqrt(1 - phi_dyn_race_x_education
* phi_dyn_race_x_education);
  beta_dyn_gender[1] <- mu_dyn_gender + beta_0_gender * sigma_dyn_gender * scale_sigma
/ sqrt(1 - phi_dyn_gender * phi_dyn_gender);
  beta_dyn_married[1] <- mu_dyn_married + beta_0_married * sigma_dyn_married *
scale_sigma / sqrt(1 - phi_dyn_married * phi_dyn_married);
  beta_dyn_party[1] <- mu_dyn_party + beta_0_party * sigma_dyn_party * scale_sigma
/ sqrt(1 - phi_dyn_party * phi_dyn_party);

  beta_fix_division[1] <- mu_fix_division * sigma_fix_division;

if (n_waves > 1) {
for (k in 2:n_waves) {
  alpha[k] <- mu_alpha;# + phi_alpha_dyn * (alpha[k-1] - mu_alpha) + alpha_raw[k-1]
```

```
* sigma_alpha_dyn * scale_sigma;

    beta_dyn_race_x_education[k] <- mu_dyn_race_x_education + phi_dyn_race_x_education
* (beta_dyn_race_x_education[k-1] - mu_dyn_race_x_education) + beta_raw_race_x_education
* sigma_dyn_race_x_education * scale_sigma;
    beta_dyn_gender[k] <- mu_dyn_gender + phi_dyn_gender * (beta_dyn_gender[k-1]
- mu_dyn_gender) + beta_raw_gender[k-1] * sigma_dyn_gender * scale_sigma;
    beta_dyn_married[k] <- mu_dyn_married + phi_dyn_married * (beta_dyn_married[k-1]
- mu_dyn_married) + beta_raw_married[k-1] * sigma_dyn_married * scale_sigma;
    beta_dyn_party[k] <- mu_dyn_party + phi_dyn_party * (beta_dyn_party[k-1] -
mu_dyn_party) + beta_raw_party[k-1] * sigma_dyn_party * scale_sigma;

    beta_fix_division[k] <- mu_fix_division * sigma_fix_division;

}
}
}

model {

  beta_0_race_x_education ~ double_exponential( 0, 1);
  beta_0_gender ~ double_exponential( 0, 1);
  beta_0_married ~ double_exponential( 0, 1);
  beta_0_party ~ double_exponential( 0, 1);

if (n_waves > 1) {
```

```
  for (k in 1:(n_waves-1)) {
    beta_raw_race_x_education[k] ~ double_exponential( 0, 1);
    beta_raw_gender[k] ~ double_exponential( 0, 1);
    beta_raw_married[k] ~ double_exponential( 0, 1);
    beta_raw_party[k] ~ double_exponential( 0, 1);
  }
}

scale_sigma ~ normal(0, 0.1);

  mu_dyn_race_x_education ~ normal(0, 0.5);
  mu_dyn_gender ~ normal(0, 0.5);
  mu_dyn_married ~ normal(0, 0.5);
  mu_dyn_party ~ normal(0, 0.5);

  mu_fix_division ~ normal(0, 1);

  sigma_dyn_race_x_education ~ normal(0, 1);
  sigma_dyn_gender ~ normal(0, 1);
  sigma_dyn_married ~ normal(0, 1);
  sigma_dyn_party ~ normal(0, 1);

  sigma_fix_division ~ normal(0, 0.5);

  phi_dyn_race_x_education ~ beta(10, 1);
  phi_dyn_gender ~ beta(10, 1);
```

```
  phi_dyn_married ~ beta(10, 1);
  phi_dyn_party ~ beta(10, 1);

  alpha_Obama_voteshare ~ normal(0, 1);
  alpha_meas_error_party_id ~ normal(0, 1);

  for(s in 1:n_waves){

    y[wave_start[s]:wave_end[s]] ~ binomial_logit(
      total[wave_start[s]:wave_end[s]],
      alpha[s]
      + beta_dyn_race_x_education[s, X_dyn[wave_start[s]:wave_end[s],1]]
      + beta_dyn_gender[s, X_dyn[wave_start[s]:wave_end[s],2]]
      + beta_dyn_married[s, X_dyn[wave_start[s]:wave_end[s],3]]
      + beta_dyn_party[s, X_dyn[wave_start[s]:wave_end[s],4]]
      + beta_fix_division[s, X_fix[wave_start[s]:wave_end[s],1]]
      + alpha_Obama_voteshare * X_cov[wave_start[s]:wave_end[s],1]
      + alpha_meas_error_party_id * X_cov[wave_start[s]:wave_end[s],2]
    );
  }
}
```