Forecasting Elections with Non-Representative Polls

Wei Wang[a], David Rothschild[b], Sharad Goel[b], Andrew Gelman[a,c]

[a]*Department of Statistics, Columbia University, New York, NY, USA*
[b]*Microsoft Research, New York, NY, USA*
[c]*Department of Political Science, Columbia University, New York, NY, USA*

Abstract

Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked for whom they intend to vote. While representative polling has historically proven to be quite effective, it comes at considerable financial and time costs. Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and often faster and at less expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After adjusting the Xbox responses via multilevel regression and poststratification, we obtain estimates in line with forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle. We conclude by arguing that non-representative polling shows promise not only for election forecasting, but also for measuring public opinion on a broad range of social, economic and cultural issues.

Keywords:
Non-representative polling, multilevel regression and poststratification, election forecasting

1. Introduction

At the heart of modern opinion polling is representative sampling, built around the goal that every individual in a particular target population such as registered or likely U.S. voters has the same probability of being sampled. From address-based, in-home interview sampling in the 1930s to random digit dialing after the growth of landlines and cellphones, leading polling organizations have put immense effort into obtaining representative samples.

Email addresses: ww2243@columbia.edu (Wei Wang), davidmr@microsoft.com (David Rothschild), sharadg@microsoft.com (Sharad Goel), gelman@stat.columbia.edu (Andrew Gelman)

We thank Jon Krosnick for helpful comments and the National Science Foundation for partial support of this research.

The wide-scale adoption of representative polling can largely be traced to a pivotal polling mishap in the 1936 U.S. presidential election campaign. During that campaign, the popular magazine *Literary Digest* conducted a mail-in survey that attracted over two million responses, a huge sample even by modern standards. The magazine, however, incorrectly predicted a landslide victory for Republican candidate Alf Landon over the incumbent Franklin Roosevelt. Roosevelt, in fact, decisively won the election, carrying every state except for Maine and Vermont. As pollsters and academics have since pointed out, the magazine's pool of respondents was highly biased: it consisted mostly of auto and telephone owners as well as the magazine's own subscribers, which underrepresented Roosevelt's core constituencies (Squire, 1988). During that same campaign, pioneering pollsters, including George Gallup, Archibald Crossley, and Elmo Roper, used considerably smaller but representative samples to predict the election outcome with reasonable accuracy (Gosnell, 1937). Accordingly, non-representative or "convenience sampling" rapidly fell out of favor with polling experts.

So why do we revisit this seemingly long-settled case? Two recent trends spur our investigation. First, random digit dialing (RDD), the standard method in modern representative polling, has suffered increasingly high non-response rates, both due to the general public's growing reluctance to answer phone surveys, and expanding technical means to screen unsolicited calls (Keeter et al., 2006). By one measure, RDD response rates have decreased from 36% in 1997 to 9% in 2012 (Kohut et al., 2012), and other studies confirm this trend (Holbrook et al., 2007; Steeh et al., 2001; Council, 2013). Assuming the initial pool of targets is representative, with such low response rates, those who ultimately answer the phone and elect to respond might not be. Even if the selection issues are not yet a serious issue for accuracy, as some have argued (Holbrook et al., 2007), the downward trend in response rates suggest an increasing need for post-sampling adjustment, and indeed the adjustment methods we present here should work just as well for surveys obtained by probability sampling as with convenience samples. The second trend driving our research is that with recent technological innovations, it is increasingly convenient and cost-effective to collect large numbers of highly non-representative samples via online surveys. What took several months for the *Literary Digest* editors to collect in 1936 can now take only a few days and for some surveys can cost just pennies per response. The challenge is to extract meaningful signal from these unconventional samples.

In this paper, we show that with proper statistical adjustment, non-representative polls are able to yield accurate presidential election forecasts, on par with those based on traditional representative polls. We proceed as follows. Section 2 describes the election survey that we conducted on the Xbox gaming platform during the 45 days leading up to the 2012 U.S. presidential race. Our Xbox sample is highly biased in two key demographic dimensions, gender and age, and the raw responses accordingly disagree with the actual outcomes. The statistical techniques we use to adjust the raw estimates are introduced in two stages. In Section 3, we construct daily estimates of voter intent via multilevel regression and post-stratification (MRP). The central idea of MRP is to partition the data into thousands of demographic cells, estimate voter intent at the cell level with a multilevel regression model,

and finally to aggregate cell-level estimates in accordance with the target population's demographic composition. Recent work suggests that non-probability samples provide worse estimates compared to probability samples (Yeager et al., 2011), but that work used simple adjustment techniques, not MRP. Even after getting good estimates of daily voter intent, however, more needs to be done to translate to election-day forecasts. Section 4 describes how to transform voter intent to projections of vote share and electoral votes. We conclude in Section 5 by discussing the potential for non-representative polling in other domains.

2. Xbox data

Our analysis is based on an opt-in poll continuously available on the Xbox gaming platform during the 45 days preceding the 2012 U.S. presidential election. Each day, three to five questions were posted, one of which gauged voter intention with the standard query, "If the election were held today, who would you vote for?" Full details of the questionnaire are in the Appendix. Respondents were allowed to answer at most once per day. The first time they participated in an Xbox poll, respondents were additionally asked to provide basic demographic information about themselves, including their sex, race, age, education, state, party ID, political ideology, and for whom they voted in the 2008 presidential election. In total, 750,148 interviews were conducted with 345,858 unique respondents—over 30,000 of whom completed five or more polls—making this one of the largest ever election panel studies.

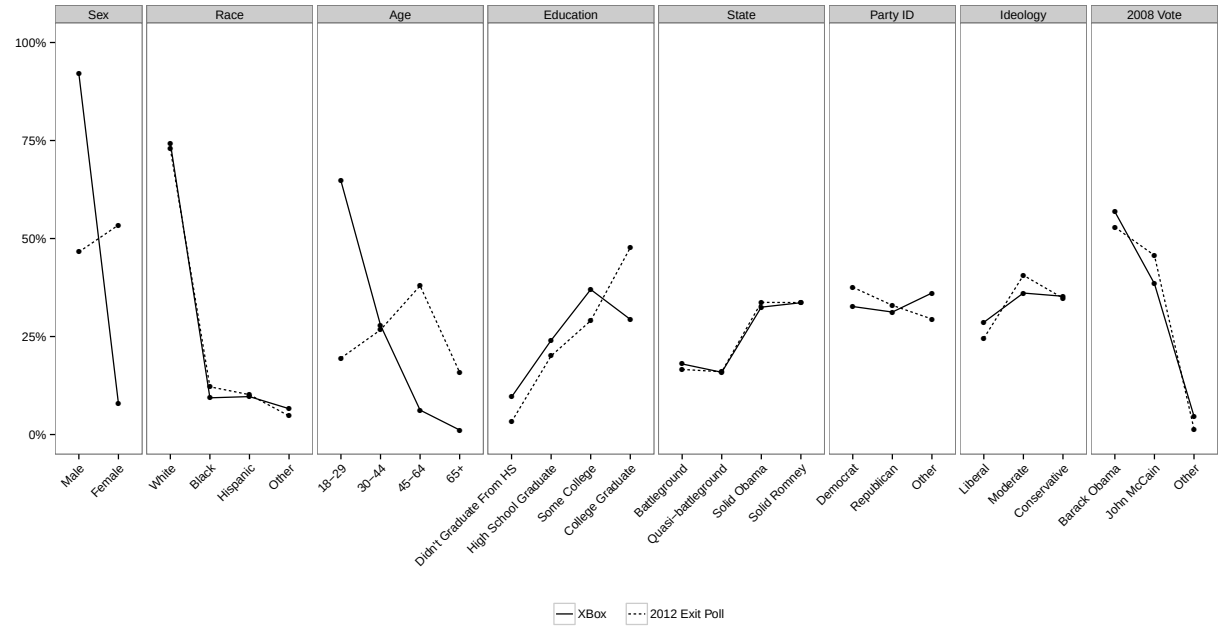

Figure 1: A comparison of the demographic, partisan, and 2008 vote distribution in the Xbox dataset and the 2012 electorate (as measured by adjusted exit polls). The sex and age distributions, as one might expect, exhibit considerable differences.

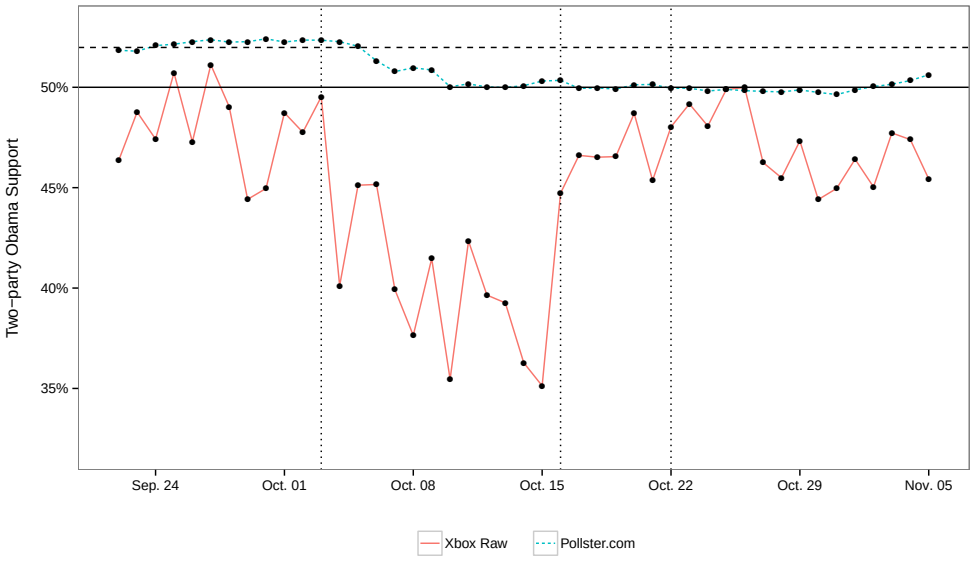

Figure 2: Daily (unadjusted) Xbox estimates of two-party Obama support during the 45 days leading up to the 2012 presidential election, which suggest a landslide victory for Mitt Romney. The dotted blue line indicates a consensus average of traditional polls (the daily aggregated polling results from Pollster.com), the horizontal dashed line at 52% indicates the actual two-party vote share obtained by Barack Obama, and the vertical dotted lines give the dates of the three presidential debates.

Despite the large sample size, the pool of Xbox respondents is far from representative of the voting population. Figure 1 compares the demographic composition of the Xbox participants to that of the general electorate, as estimated via the 2012 national exit poll.[2] The most striking differences are for age and sex. As one might expect, young men dominate the Xbox population: 18-to-29-year-olds comprise 65% of the Xbox dataset, compared to 19% in the exit poll; and men make up 93% of the Xbox sample but only 47% of the electorate. Political scientists have long observed that both age and sex are strongly correlated with voting preferences (Kaufmann and Petrocik, 1999), and indeed these discrepancies are apparent in the unadjusted time-series of Xbox voter intent shown in Figure 2. In contrast to estimates

[2]For ease of interpretation, in Figure 1 we group states into 4 categories: (1) battleground states (Colorado, Florida, Iowa, New Hampshire, Ohio, and Virginia), the five states with the highest amounts of TV spending plus New Hampshire, which had the highest per-capita spending; (2) quasi-battleground states (Michigan, Minnesota, North Carolina, Nevada, New Mexico, Pennsylvania, and Wisconsin), which round out the states where the campaigns and their affiliates made major TV buys; (3) solid Obama states (California, Connecticut, District of Columbia, Delaware, Hawaii, Illinois, Maine, Maryland, Massachusetts, New Jersey, New York, Oregon, Rhode Island, Vermont, and Washington); and (4) solid Romney states (Alabama, Alaska, Arizona, Arkansas, Georgia, Idaho, Indiana, Kansas, Kentucky, Louisiana, Mississippi, Missouri, Montana, Nebraska, North Dakota, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, Utah, West Virginia, and Wyoming).

based on traditional, representative polls (indicated by the dotted blue line in Figure 2), the uncorrected Xbox sample suggests a landslide victory for Mitt Romney, reminiscent of the infamous *Literary Digest* error.

3. Estimating voter intent with multilevel regression and poststratification

3.1. Multilevel regression and poststratification

To transform the raw Xbox data into accurate estimates of voter intent in the general electorate, we make use of the rich demographic information that respondents provide. In particular we *poststratify* the raw Xbox responses to mimic a representative sample of likely voters. Poststratification is a popular method for correcting for known differences between sample and target populations (Little, 1993). The core idea is to partition the population into cells based on combinations of various demographic and political attributes, use the sample to estimate the response variable within each cell, and finally to aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population. Using y to indicate the outcome of interest, the poststratification estimate is defined by,

$$\hat{y}^{\text{PS}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population. We can analogously derive an estimate of y at any subpopulation level s (e.g., voter intent in a particular state) by

$$\hat{y}_s^{\text{PS}} = \frac{\sum_{j \in J_s} N_j \hat{y}_j}{\sum_{j \in J_s} N_j}$$

where J_s is the set of all cells that comprise s. As is readily apparent from the form of the poststratification estimator, the key is to obtain accurate cell-level estimates, as well as estimates for the cell sizes.

One of the most common ways to generate cell-level estimates is to simply average sample responses within each cell. If we assume that within a cell the sample is drawn at random from the larger population, this yields an unbiased estimate. However, this assumption of cell-level simple random sampling is only reasonable when the partition is sufficiently fine; on the other hand, as the partition becomes finer, the cells become sparse, and the empirical sample averages become unstable. We address these issues by instead generating cell-level estimates via a regularized regression model, namely multilevel regression. This combined model-based poststratification strategy, known as multilevel regression and poststratification (MRP), has been used to obtain accurate small-area subgroup estimates, such as for public opinion and voter turnout in individual states and demographic subgroups (Park et al., 2004; Lax and Phillips, 2009; Ghitza and Gelman, 2013).

More formally, applying MRP in our setting comprises two steps. First a Bayesian hierarchical model is fit to obtain estimates for sparse poststratification cells; second, one averages over the cells, weighting by a measure of forecasted voter turnout, to get state and national-level estimates. Specifically, we generate the cells by considering all possible combinations

of sex (2 categories), race (4 categories), age (4 categories), education (4 categories), state (51 categories), party ID (3 categories), ideology (3 categories) and 2008 vote (3 categories), which partition the data into 176,256 cells.[3] Including the political variables is important because they are strong predictors of vote intention. Poststratifying on party identification has sometimes been controversial (Pollster) but we are comfortable with it here, first because it tends to vary more slowly than vote intentions and political attitudes (Gelman and King, 1993; Cavan Reilly and Katz, 2001), and second because party identification and the other background variables in the Xbox survey were measured only once during the campaign, at the time of the respondents entry into the panel.

We fit two, nested multilevel logistic regressions to estimate candidate support in each cell. The first of the two models predicts whether a respondent supports a major-party candidate (i.e., Obama or Romney), and the second predicts support for Obama given that the respondent supports a major-party candidate. Following the notation of Gelman and Hill (2007), the first model is given by

$$\begin{aligned}\Pr(Y_i \in \{\text{Obama, Romney}\}) = \\ \text{logit}^{-1}\big(&\alpha_0 + \alpha_1(\text{state last vote share}) \\ &+ a^{\text{state}}_{j[i]} + a^{\text{edu}}_{j[i]} + a^{\text{sex}}_{j[i]} + a^{\text{age}}_{j[i]} + a^{\text{race}}_{j[i]} + a^{\text{party ID}}_{j[i]} + b^{\text{ideology}}_{j[i]} + b^{\text{last vote}}_{j[i]}\big)\end{aligned} \tag{1}$$

where α_0 is the fixed baseline intercept, and α_1 is the fixed slope for Obama's fraction of two-party vote share in the respondent's state in the last presidential election. The terms $a^{\text{state}}_{j[i]}$, $a^{\text{edu}}_{j[i]}$, $a^{\text{sex}}_{j[i]}$ and so on—which in general we denote by $a^{\text{var}}_{j[i]}$—correspond to varying coefficients associated with each categorical variable. Here the subscript $j[i]$ indicates the cell to which the i-th respondent belongs. For example, $a^{\text{age}}_{j[i]}$ takes values from $\{a^{\text{age}}_{18-29}, a^{\text{age}}_{30-44}, a^{\text{age}}_{45-64}, a^{\text{age}}_{65+}\}$ depending on the cell membership of the i-th respondent. The varying coefficients $a^{\text{var}}_{j[i]}$ are given independent prior distributions

$$a^{\text{var}}_{j[i]} \sim N(0, \sigma^2_{\text{var}}).$$

To complete the full Bayesian specification, the variance parameters are assigned a hyperprior distribution

$$\sigma^2_{\text{var}} \sim \text{inv-}\chi^2(\nu, \sigma^2_0),$$

with a weak prior specification for the remaining parameters, ν and σ_0. The benefit of using a multilevel model is that estimates for relatively sparse cells can be improved through "borrowing strength" from demographically similar cells that have richer data. Similarly,

[3] All demographic variables are collected prior to respondents' first poll, alleviating concerns that respondents may adjust their demographic responses to be inline with their voter intention (e.g., a new Obama supporter switching his or her party ID from Republican to Democrat).

the second model is defined by

$$\begin{aligned}\Pr(Y_i = \text{Obama} \,|\, Y_i \in \{\text{Obama, Romney}\}) = \\ \text{logit}^{-1}\big(\beta_0 + \beta_1(\text{state last vote share}) \\ + b^{\text{state}}_{j[i]} + b^{\text{edu}}_{j[i]} + b^{\text{sex}}_{j[i]} + b^{\text{age}}_{j[i]} + b^{\text{race}}_{j[i]} + b^{\text{party ID}}_{j[i]} + b^{\text{ideology}}_{j[i]} + b^{\text{last vote}}_{j[i]}\big)\end{aligned} \tag{2}$$

and

$$\begin{aligned} b^{\text{var}}_{j[i]} &\sim N(0, \eta^2_{\text{var}}), \\ \eta^2_{\text{var}} &\sim \text{inv-}\chi^2(\mu, \eta^2_0). \end{aligned}$$

Jointly, Eqs. (1) and (2) define a Bayesian model that describes the data. Ideally, we would perform a fully Bayesian analysis to obtain the posterior distribution of the parameters. However, for computational convenience, we use the approximate marginal maximum likelihood estimates obtained from the `glmer()` function in the `R` package `lme4` (Bates et al., 2013). We run the multilevel model daily using a four-day moving window, aggregating the data collected on that day and the previous three days, to make cell-level estimates for each of the 45 days leading up to the election.

Having detailed the multilevel regression step, we now turn to poststratification, where cell-level estimates are weighted by the proportion of the electorate in each cell and aggregated to the appropriate level (e.g., state or national). To compute cell weights, we require cross-tabulated population data. One commonly used source for such data is the Current Population Survey (CPS); however, the CPS does not include some key poststratification variables, such as party identification. We thus instead use exit poll data from the 2008 presidential election. Exit polls are conducted on election day outside voting stations to record the choices of exiting voters, and they are generally used by researchers and news media to analyze the demographic breakdown of the vote (after a post-election adjustment that aligns the weighted responses to the reported state-by-state election results). In total, 101,638 respondents were surveyed in the state and national exit polls. We use the exit poll from 2008, not 2012, because this means that in theory our method as described here could have been used to generate real-time predictions during the 2012 election campaign. Admittedly, this approach puts our prediction at a disadvantage since we cannot capture the demographic shifts of the intervening four years. While combining exit poll and CPS data would arguably yield improved results, for simplicity and transparency we exclusively use the 2008 exit poll summaries for our poststratification.

3.2. National and state voter intent

Figure 3 shows the adjusted two-party Obama support for the last 45 days of the election. Compared with the uncorrected estimates in Figure 2, the MRP-adjusted estimates yield a much more reasonable timeline of Obama's standing over the course of the final weeks of the campaign. With a clear advantage at the beginning, Obama's support slipped rapidly after the first presidential debate—though never falling below 50%—and gradually recovered, building up a decisive lead in the final days.

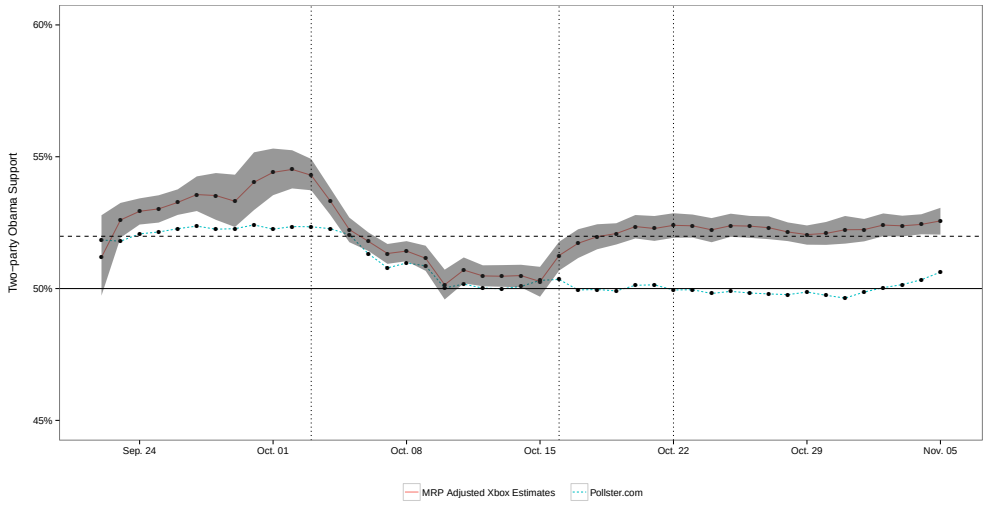

Figure 3: National MRP-adjusted voter intent of two-party Obama support over the 45-day period and the associated 95% confidence bands. The horizontal dashed line indicates the actual two-party Obama vote share. The three vertical dotted lines indicate the presidential debates. Compared with the raw responses in Figure 2, the MRP-adjusted voter intent is much more reasonable, and voter intent in the last few days is close to the actual outcome. For comparison, the daily aggregated polling results from Pollster.com, shown as the blue dotted line, are further away from the actual vote share than the estimates generated from the Xbox data in the last few days.

On the day before the election, our estimate of voter intent is off by a mere 0.6 percentage points from the actual outcome (indicated by the dotted horizontal line). Voter intent in the weeks prior to the election does not directly equate to an estimate of vote share on election day—a point we return to in Section 4. As such, it is difficult to evaluate the accuracy of our full time-series of estimates. Our estimates are not only intuitively reasonable and are also in line with prevailing estimates based on traditional, representative polls. In particular, our estimates roughly track—and are even arguably better than—those from Pollster.com, one of the leading poll aggregators during the 2012 campaign. We are following now-standard practice by comparing to an aggregate of nationally reported polls, not to any selected subset of polls that might be judged to be of highest quality. Again, our goal here is not to bash conventional polling but rather to demonstrate how, with sophisticated adjustments, we can match that performance even with a highly non-representative opt-in sample.

National vote share receives considerable media attention, but state-level estimates are particularly relevant for many stakeholders given the role of the Electoral College in selecting the winner (Rothschild, 2013). Forecasting state-by-state races is a challenging problem due to the interdependencies in state outcomes, the logistical difficulties of measuring state-level vote preference, and the effort required to combine information from various sources (Lock and Gelman, 2010). The MRP framework, however, provides a straightforward methodology for generating state-level results. Namely, we use the same cell-level estimates employed in

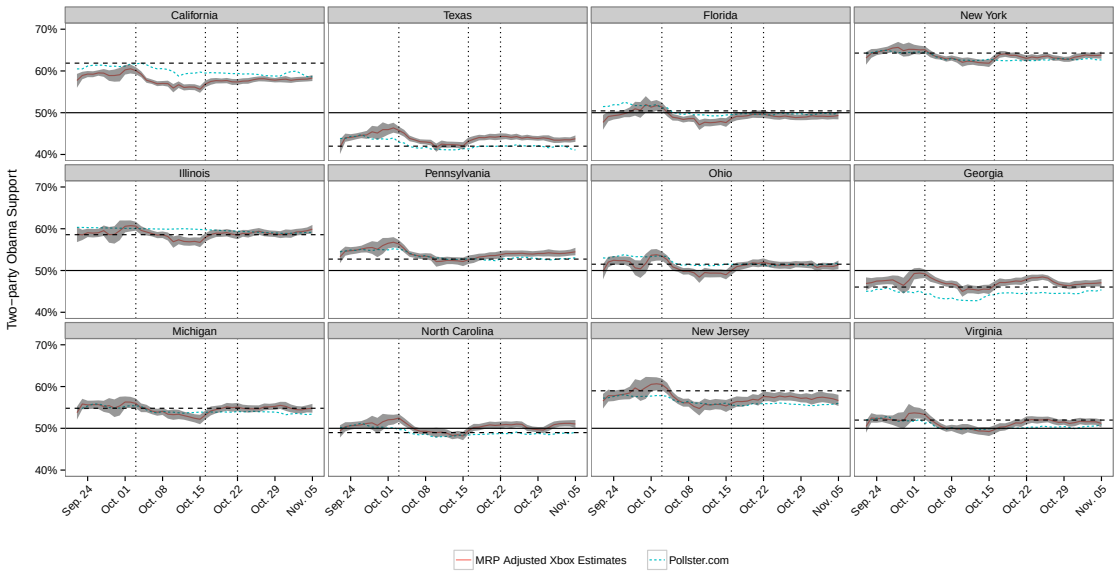

Figure 4: MRP-adjusted daily voter intent for the 12 states with the most electoral votes, and the associated 95% confidence bands. The horizontal dashed lines in each panel give the actual two-party Obama vote shares in that state. The mean and median absolute errors of the last day voter intent across the 51 Electoral College races are 2.5 and 1.8 percentage points, respectively. The state-by-state daily aggregated polling results from Pollster.com, given in the dotted blue lines, are broadly consistent with the estimates from the Xbox data.

the national estimate, as generated via the multilevel model in Eqs. (1) and (2), and we then poststratify to each state's demographic composition. In this manner, the Xbox responses can be used to construct estimates of voter intent over the last 45 days of the campaign for all 51 Electoral College races.

Figure 4 shows two-party Obama support for the 12 states with the most electoral votes. The state timelines share similar trends (e.g., support for Obama dropping after the first debate), but also have their own idiosyncratic movements, an indication of a reasonable blend of national and state-level signals. To demonstrate the accuracy of the MRP-adjusted estimates, we plot, in dotted blue lines in Figure 4, the estimates generated by Pollster.com, which are broadly consistent with our state-level MRP estimates. Moreover, across the 51 Electoral College races, the mean and median absolute errors of our estimates on the day before the election are just 2.5 and 1.8 percentage points, respectively.

3.3. Voter intent for demographic subgroups

Apart from Electoral College races, election forecasting often focuses on candidate preference among demographic subpopulations. Such forecasts are of significant importance in modern political campaigns, which often employ targeted campaign strategies (Hillygus and Shields, 2009). In the highly non-representative Xbox survey, certain subpopulations are heavily underrepresented and plausibly suffer from strong self-selection problems. This begs

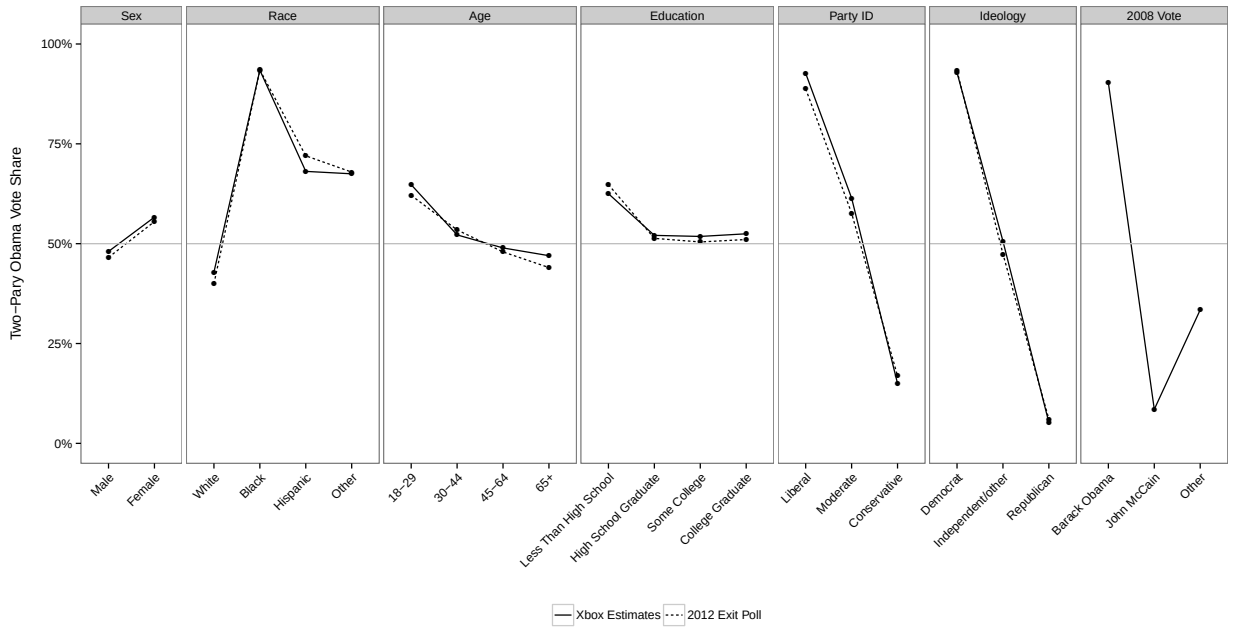

Figure 5: Comparison of two-party Obama vote share for various demographic subgroups, as estimated from the 2012 national exit poll and from the Xbox data on the day before the election.

the question, can we reasonably expect to estimate the views of older women on a platform that largely caters to young men?

It is straightforward in MRP to estimate voter intent among any collection of demographic cells: we again use the same cell-level estimates as in the national and state settings, but poststratify to the desired target population. For example, to estimate voter intent among women, the poststratification weights are based on the relative number of women in each demographic cell. To illustrate this approach, we compute Xbox estimates of Obama support for each level of our categorical variables (e.g., males, females, Whites, Blacks, etc.) on the day before the election, and compare those with the actual voting behavior of those same groups as estimated by the 2012 national exit poll. As seen in Figure 5, the Xbox estimates are remarkably accurate, with a median absolute difference of 1.5 percentage points between the Xbox and the exit poll numbers.[4]

Not only do the Xbox data facilitate accurate estimation of voter intent across these single-dimensional demographic categories, but they also do surprisingly well at estimating two-way interactions (e.g., candidate support among 18–29 year-old Hispanics, and liberal college graduates). Figure 6 shows this result, plotting the Xbox estimates against those derived from the exit polling data for each of the 149 two-dimensional demographic subgroups.[5]

[4]Respondents' 2008 vote was not asked on the 2012 exit poll, so we exclude that comparison from Figure 5.

[5]State contestedness is excluded from the two-way interaction groups since the 2012 state exit polls are

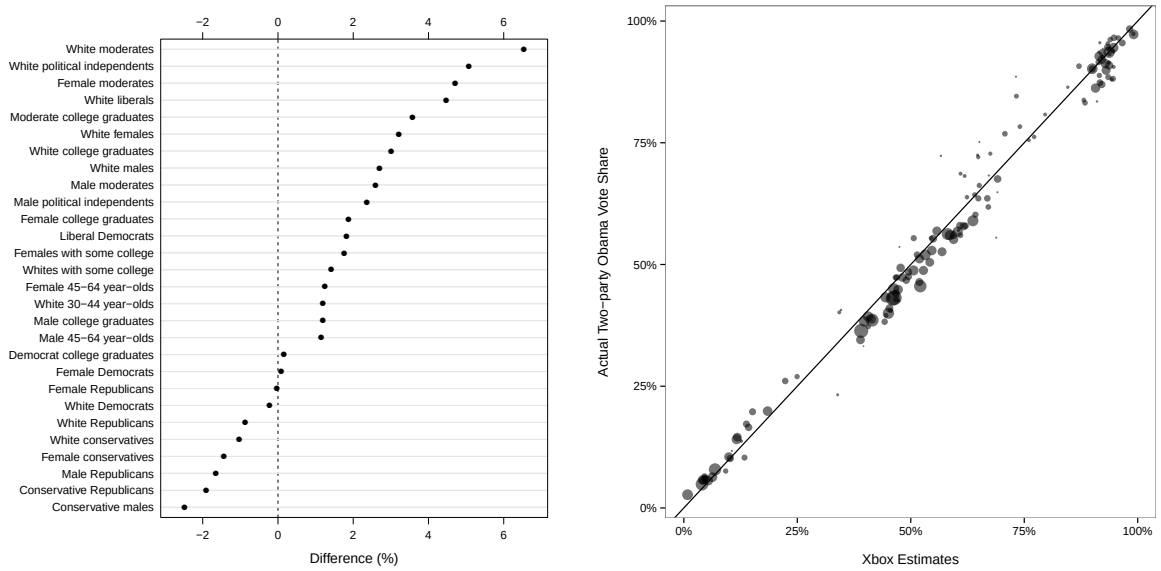

Figure 6: Left panel: Differences between the Xbox MRP-adjusted estimates and the exit poll estimates for the 30 largest two-dimensional demographic subgroups, ordered by the difference. Positive values indicate the Xbox estimate is larger than the corresponding exit poll estimate. Among these 30 subgroups, the median and mean absolute differences are 1.9 and 2.2 percentage points, respectively. Right panel: Two-party Obama support as estimated from the 2012 national exit poll and from the Xbox data on the day before the election, for various two-way interaction demographic subgroups (e.g., 65+ year-old women). The sizes of the dots are proportional to the population sizes of the corresponding subgroups.

Most points lie close to the diagonal, indicating that the Xbox and exit poll estimates are in agreement. Specifically, for women who are 65 and older—a group whose preferences one might a priori believe are hard to estimate from the Xbox data—the difference between Xbox and the exit poll is a mere one percentage point (49.5% and 48.5%, respectively). Across all the two-way interaction groups, the median absolute difference is just 2.4 percentage points. As indicated by the size of the points in Figure 6, the largest differences occur for relatively small demographic subgroups (e.g., liberal Republicans), for which both the Xbox and exit poll estimates are less reliable. For the 30 largest demographic subgroups, Figure 6 lists the differences between Xbox and exit poll estimates. Among these largest subgroups, the median absolute difference drops to just 1.9 percentage points.

not yet available, and the 2012 national exit poll does not have enough data to reliably estimate state interactions; 2008 vote is also excluded, as it was not asked in the 2012 exit poll. The "other" race category was also dropped as it was not consistently defined across the Xbox and exit poll datasets.

4. Forecasting election day outcomes

4.1. Converting voter intent to forecasts

As mentioned above, daily estimates of voter intent do not directly correspond to estimates of vote share on election day. There are two key factors for this deviation. First, opinion polls (both representative and non-representative ones) only gauge voter preference on the particular day when the poll is conducted, with the question typically phrased as, "if the election were held today." Political scientists and pollsters have long observed that such stated preferences are prone to several biases, including the anti-incumbency bias, in which the incumbent's polling numbers tend to be lower than the ultimate outcome (Campbell, 2008; Erikson and Wlezien, 2008; Rothschild, 2013), and the fading early lead bias, in which a big lead early in the campaign tends to diminish as the election gets closer (Erikson and Wlezien, 2008). Moreover, voters' attitudes are affected by information revealed over the course of the campaign, so preferences weeks or months before election day are at best a noisy indicator of one's eventual vote. Second, estimates of vote share require a model of likely voters. That is, opinion polls measure preferences among a hypothetical voter pool, and are thus accurate only to the extent that this pool captures those who actually turn out to vote on election day. Both of these factors introduce significant complications in forecasting election day outcomes.

To convert daily estimates of voter intent to election day predictions—which we hereafter refer to as *calibrating* voter intent—we compare daily voter intent in previous elections to the ultimate outcomes in those elections. Specifically, we collected historical data from three previous U.S. presidential elections, in 2000, 2004, and 2008. For each year, we obtained top-line (i.e., not individual-level) national and state estimates of voter intent from all available polls conducted in those elections.[6] From this collection of polling data, we then constructed daily estimates of voter intent by taking a moving average of the poll numbers, in a similar manner to the major poll aggregators. We rely on traditional, representative polls to reconstruct historical voter intent; in principle, however, we could have started with non-representative polls if such data were available in previous election cycles.

We next infer a mapping from voter intent to election outcomes by regressing election day vote share on the historical time-series of voter intent. The key difference between our approach and previous related work (Erikson and Wlezien, 2008; Rothschild, 2009) is that we explicitly model state-level correlations, via nested national and state models and correlated error terms. Specifically, we first fit a national model given by

$$y_e^{\text{US}} = a_0 + a_1 x_{t,e}^{\text{US}} + a_2 |x_{t,e}^{\text{US}}| x_{t,e}^{\text{US}} + a_3 t x_{t,e}^{\text{US}} + \eta(t,e)$$

where y_e^{US} is the national election day vote share of the incumbent party candidate in election year e, $x_{t,e}^{\text{US}}$ is the national voter intent of the incumbent party candidate at t days before the election in year e, and $\eta \sim N(0, \sigma^2)$ is the error term. Both y_e^{US} and $x_{t,e}^{\text{US}}$ are offset by 0.5, so the values run from -0.5 to 0.5 rather than 0 to 1. The term involving the absolute value

[6] We collected the polling data from `Pollster.com` and `RealClearPolitics.com`.

of voter intent pulls the vote share prediction toward 50%, capturing the diminishing early lead effect. We do not include a main effect for time since it seems unlikely that the number of days until the election itself contributes to the final vote share directly, but rather time contributes through its interaction with the voter intent (which we do include in the model).

Similarly, the state model is given by

$$y^{\text{ST}}_{s,e} = b_0 + b_1 x^{\text{ST}}_{s,t,e} + b_2 |x^{\text{ST}}_{s,t,e}| x^{\text{ST}}_{s,t,e} + b_3 t x^{\text{ST}}_{s,t,e} + \varepsilon(s,t,e)$$

where $y^{\text{ST}}_{s,e}$ is the election day state vote share of the state's incumbent party candidate[7] at day t, $x^{\text{ST}}_{s,t,e}$ is the state voter intent at day t, and ϵ is the error term. The outcome $y^{\text{ST}}_{s,e}$ is offset by the national projected vote share on that day as fit with the national calibration model, and $x^{\text{ST}}_{s,t,e}$ is offset by that day's national voter intent. Furthermore, we impose two restrictions on the magnitude and correlation structure of the error term $\varepsilon(s,t,e)$. First, since the uncertainty naturally decreases as the election gets closer (as t becomes smaller), we apply the heteroscedastic structure $\text{Var}(\varepsilon(s,t,e)) = (t+a)^2$, where a is a constant to be estimated from the data. Second, the state-specific movements within each election year are allowed to be correlated. For simplicity, and as in Chen et al. (2008), we assume these correlations are uniform (i.e., all pairwise correlations are the same), which creates one more parameter to be estimated from the data. We fit the full calibration model with the `gls()` function in the `R` package `nlme` (Pinheiro et al., 2012).

In summary, the procedure for generating election day forecasts proceeds in three steps:

1. Estimate the joint distribution of state and national voter intent by applying MRP to the Xbox data, as described in Section 3.
2. Fit the nested calibration model described above on historical data to obtain point estimates for the parameters, including estimates for the error terms.
3. Convert the distribution of voter intent to election day forecasts via the fitted calibration model.

4.2. National and state election day forecasts

Figure 7 plots the projected vote shares and pointwise 95% confidence bands over time for the 12 states with the most electoral votes. Though these time-series look quite reasonable, it is difficult to assess their accuracy as there are no ground truth estimates to compare with in the weeks prior to the election. As a starting point, we compare our state-level estimates to those generated by prediction markets, which are widely considered to be among the most accurate sources for political predictions (Rothschild, 2013; Wolfers and Zitzewitz, 2004). For each state, prediction markets produce daily probabilities of victory. Though Figure 7 plots our forecasts in terms of expected vote share, our estimation procedure in fact yields the full distribution of outcomes, and so we can likewise convert our estimates to probabilistic forecasts. Figure 8 shows this comparison, where the prediction market

[7]State incumbent parties are defined as the state-by-state winners from the previous election, which is more meaningful in this context than simply using the national incumbent.

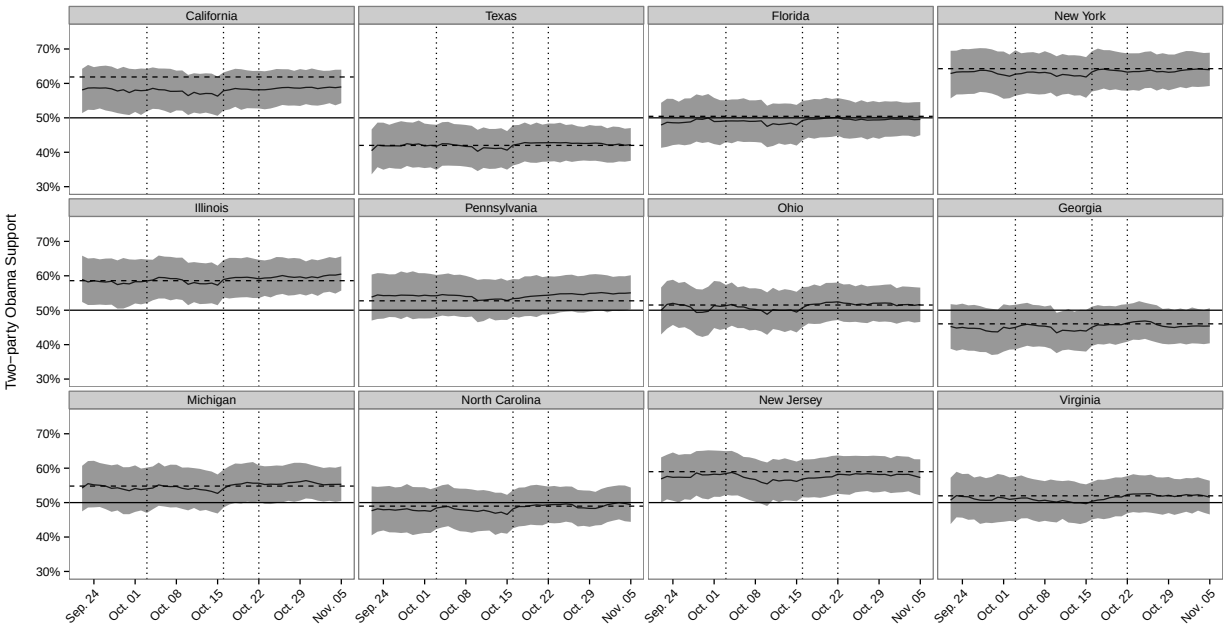

Figure 7: Projected Obama share of the two-party vote on election day for each of the 12 states with the most electoral votes, and associated 95% confidence bands. Compared to the MRP-adjusted voter intent in Figure 4, the projected two-party Obama support is more stable, and the North Carolina race switches direction after applying the calibration model. Additionally, the confidence bands become much wider and give more reasonable state-by-state probabilities of Obama victories.

estimate is derived by averaging the two largest election markets, Betfair and Intrade. Our probabilistic estimates are largely consistent with the prediction market probabilities. In fact, for races with little uncertainty such as in Texas and Massachusetts, the Xbox estimates do not seem to suffer from the longshot bias common to prediction markets (Rothschild, 2009), and instead yield probabilities closer to 0 or 1. For tighter races, the Xbox estimates—although still highly correlated with the prediction market probabilities—look more volatile, especially in the early part of the 45-day period. Since the ground truth is not clearly defined, it is difficult to evaluate which method—Xbox or prediction markets—yields better results. From a Bayesian perspective, if one believes the stability shown by prediction markets, this could be incorporated into the structure of the Xbox calibration model.

With the full state-level outcome distribution, we can also estimate the distribution of Electoral College votes. Figure 9 plots the median projected electoral votes for Obama over the last 45-days of the election, together with the 95% confidence band. In particular, on the day before the election, our model estimates Obama had an 88% chance of victory, in line with estimates based on traditional polling data. For example, Obama was given a 91% chance of victory, using a method built from Jackman (2005). Zooming in on the day before the election, Figure 10 shows the full predicted distribution of electoral votes for

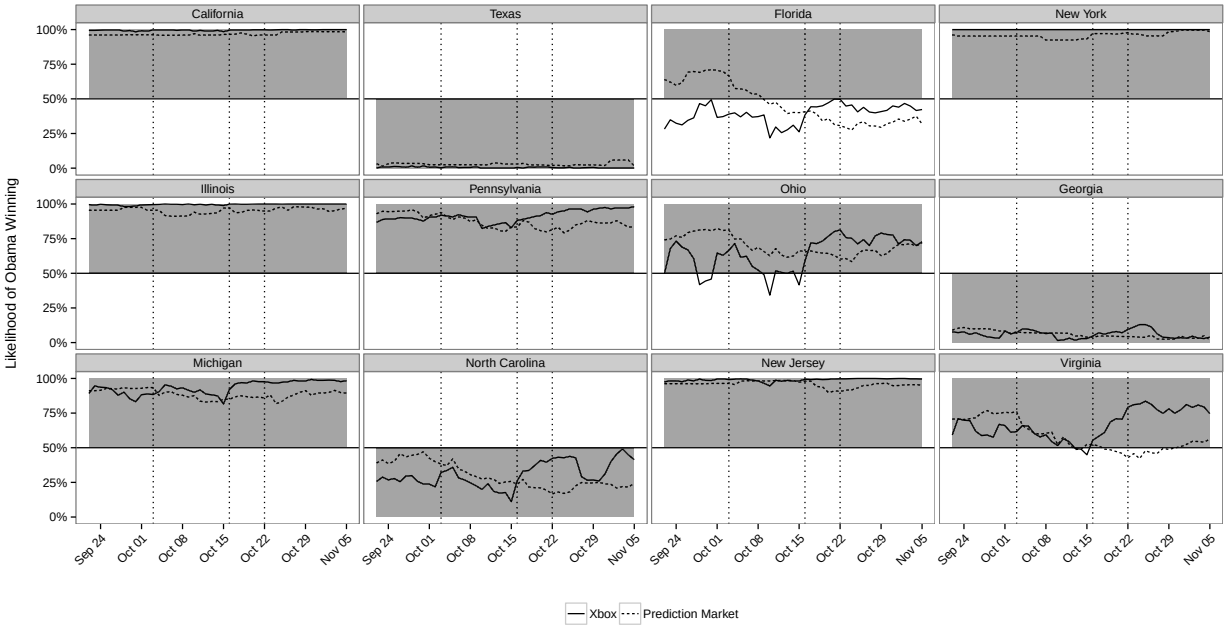

Figure 8: Comparison between the probability of Obama winning the 12 largest Electoral College races based on Xbox data and on prediction market data. The prediction market data are the average of the raw Betfair and Intrade prices from winner-take-all markets. The three vertical lines represent the dates of three presidential debates. The shaded halves indicate the direction that race went.

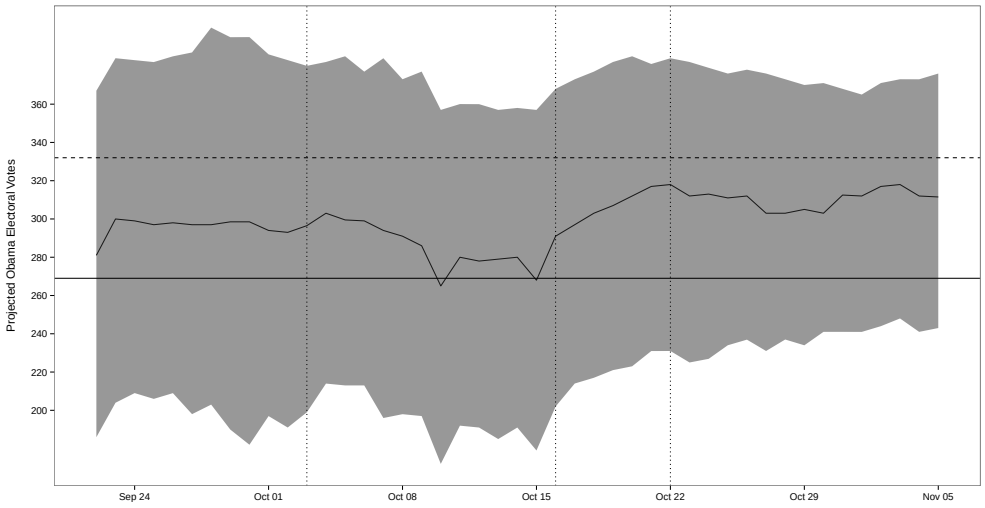

Figure 9: Daily projections of Obama electoral votes in the 45-day period leading up to the 2012 election and associated 95% confidence bands. The solid line represents the median of the daily distribution. The horizontal dashed line represents the actual electoral votes, 332, that Obama captured in 2012 election. Three vertical dotted lines indicate the dates of three presidential debates.

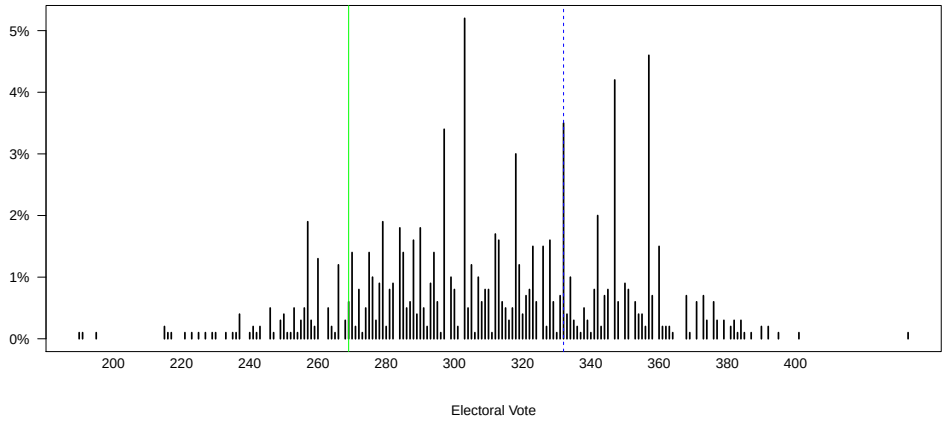

Figure 10: Projected distribution of electoral votes for Obama one day before the election. The green vertical dotted line represents 269, the minimum number of electoral votes that Obama needs for a tie. The blue vertical dashed line gives 332, the actual number of electoral votes captured by Obama. The estimated likelihood of Obama winning the electoral vote is 88%.

Obama. Compared to the actual 332 votes that Obama captured, we estimate a median of 312 votes, with the most likely outcome being 303. Though this distribution of Electoral College outcomes seems reasonable, it does appear to have higher variance than one might expect. In particular, the extreme outcomes seem to have unrealistically high likelihood of occurring, which is likely a byproduct of our calibration model not fully capturing the state-level correlation structure. Nonetheless, given that our forecasts are based on a highly biased convenience sample of respondents, the model predictions are remarkably good.

5. Conclusion

Election forecasts not only need to be accurate, but also relevant, timely, and cost-effective. In this paper, we construct forecasts satisfying all of these requirements using extremely non-representative data. Though our data were collected on a proprietary polling platform, in principle one can aggregate such non-representative samples at a fraction of the cost of conventional survey designs. Moreover, the data produce forecasts that are both relevant and timely, as they can be updated faster and more regularly than standard election polls. Thus, the key question—and one of the main contributions of this paper—is to assess the extent to which one can generate accurate predictions from non-representative samples. Since there is limited ground truth for election forecasts, definitely establishing the accuracy of our predictions is difficult. Nevertheless, we show that the MRP-adjusted and calibrated Xbox estimates are intuitively reasonable, and also similar to those generated by more traditional means.

Though our approach performed quite well, it required detailed data. In the face of insufficient demographic information on respondents, inadequate population-level statistics, or lack of historical election poll results, it would have been difficult to generate accurate forecasts from non-representative data. Further, while much of our procedure is fairly mechanical, selecting the appropriate modeling framework requires some care. Fortunately, however, at least with the Xbox data the regression estimates are stable after including only a few key demographic variables (sex, age, state, race and party identification).

The greatest impact of non-representative polling will likely not be for presidential elections, but rather for smaller, local elections and specialized survey settings, where it is impractical to deploy traditional methods due to cost and time constraints. For example, non-representative polls could be used in Congressional elections, where there are currently only sparse polling data. Non-representative polls could also supplement traditional surveys (e.g., the General Social Survey) by offering preliminary results at shorter intervals. Finally, when there is a need to identify and track pivotal events that affect public opinion, non-representative polling offers the possibility of cost-effective continuous data collection. Standard representative polling will certainly continue to be an invaluable tool for the foreseeable future. However, 75 years after the *Literary Digest* failure, non-representative polling (followed by appropriate post-data adjustment) is due for further exploration, for election forecasting and in social research more generally.

References

Bates, D., Maechler, M., Bolker, B., 2013. lme4: Linear mixed-effects models using S4 classes. URL: `http://CRAN.R-project.org/package=lme4`. r package version 0.999999-2.

Campbell, J.E., 2008. The American campaign: US presidential campaigns and the national vote. volume 6. Texas A&M University Press.

Cavan Reilly, A.G., Katz, J.N., 2001. Post-stratification without population level information on the post-stratifying variable, with application to political polling. Journal of the American Statistical Association 96, 1–11.

Chen, M.K., Ingersoll, J.E., Kaplan, E.H., 2008. Modeling a presidential prediction market. Management Science 54, 1381–1394.

Council, N.R., 2013. Nonresponse in Social Science Surveys: A Research Agenda. The National Academies Press. URL: `http://www.nap.edu/openbook.php?record_id=18293`.

Erikson, R.S., Wlezien, C., 2008. Are political markets really superior to polls as election predictors? Public Opinion Quarterly 72, 190–215.

Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

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, 409–451.

Ghitza, Y., Gelman, A., 2013. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. American Journal of Political Science 57, 762–776.

Gosnell, H.F., 1937. How accurate were the polls? Public Opinion Quarterly 1, 97–105.

Hillygus, D.S., Shields, T.G., 2009. The persuadable voter: Wedge issues in presidential campaigns. Princeton University Press.

Holbrook, A., Krosnick, J.A., Pfent, A., et al., 2007. The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. Wiley. pp. 499–528.

Jackman, S., 2005. Pooling the polls over an election campaign. Australian Journal of Political Science 40, 499–517.

Kaufmann, K.M., Petrocik, J.R., 1999. The changing politics of american men: Understanding the sources of the gender gap. American Journal of Political Science 43, 864–887.

Keeter, S., Kennedy, C., Dimock, M., Best, J., Craighill, P., 2006. Gauging the impact of growing nonresponse on estimates from a national rdd telephone survey. Public Opinion Quarterly 70, 759–779.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., Christian, L., 2012. Assessing the representativeness of public opinion surveys. Pew Research Center for The People & The Press 15, 2012.

Lax, J.R., Phillips, J.H., 2009. How should we estimate public opinion in the states? American Journal of Political Science 53, 107–121.

Little, R.J., 1993. Post-stratification: A modeler's perspective. Journal of the American Statistical Association 88, 1001–1012.

Lock, K., Gelman, A., 2010. Bayesian combination of state polls and election forecasts. Political Analysis 18, 337–348.

Park, D.K., Gelman, A., Bafumi, J., 2004. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. Political Analysis 12, 375–385.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2012. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-104.

Pollster. Should pollsters weight by party identification?, 2004. URL: `http://www.pollster.com/faq/should_pollsters_weight_by_par.php`.

Rothschild, D., 2009. Forecasting elections comparing prediction markets, polls, and their biases. Public Opinion Quarterly 73, 895–916.

Rothschild, D., 2013. Combining forecasts: Accurate, relevant, and timely. Working paper.

Squire, P., 1988. Why the 1936 literary digest poll failed. Public Opinion Quarterly 52, 125–133.

Steeh, C., Kirgis, N., Cannon, B., DeWitt, J., 2001. Are they really as bad as they seem? nonresponse rates at the end of the twentieth century. Journal of Official Statistics 17, 227–248.

Wolfers, J., Zitzewitz, E., 2004. Prediction markets. Technical Report. National Bureau of Economic Research.

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A., Wang, R., 2011. Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. Public Opinion Quarterly 75, 709–747.

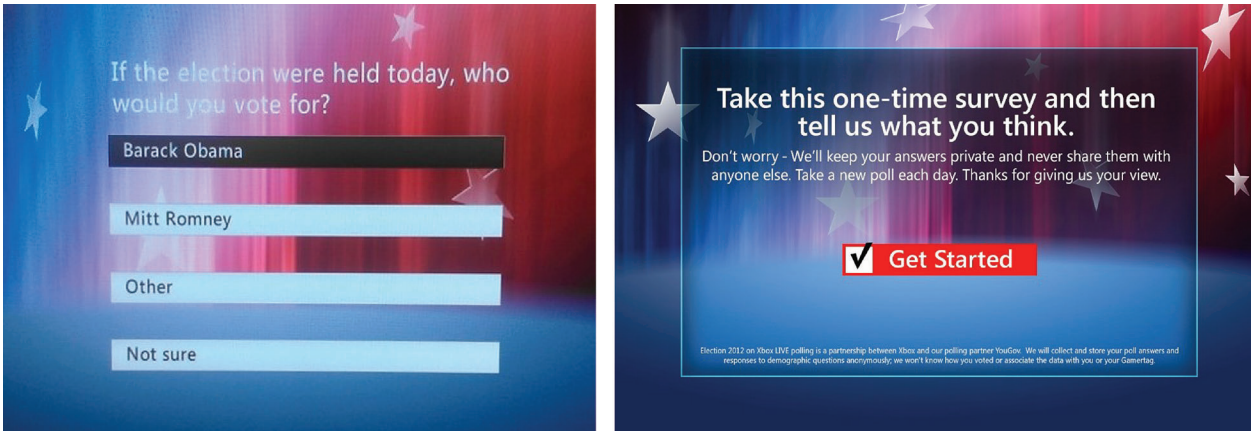

Figure A.1: The left panel shows the vote intention question, and the right panel shows what respondents were presented with during their first visit to the poll.

Appendix A. Questionnaire

The only way to answer the polling questions was via the Xbox Live gaming platform. There was no invitation or permanent link to the poll, and so respondents had to locate it daily on the Xbox Live's home page and click into it. The first time a respondent opted-into the poll, they were directed to answer the nine demographics questions listed below. On all subsequent times, respondents were immediately directed to answer between three and five daily survey questions, one of which was always the vote intention question.

Intention Question: If the election were held today, who would you vote for?
Barack Obama\Mitt Romney\Other\Not Sure

Demographics Questions:

1. Who did you vote for in the 2008 Presidential election?
   Barack Obama\John McCain\Other candidate\Did not vote in 2008
2. Thinking about politics these days, how would you describe your own political viewpoint?
   Liberal\Moderate\Conservative\Not sure
3. Generally speaking, do you think of yourself as a ...?
   Democrat\Republican\Independent\Other
4. Are you currently registered to vote?
   Yes\No\Not sure
5. Are you male or female?
   Male\Female
6. What is the highest level of education that you have completed?
   Did not graduate from high school\High school graduate\Some college or 2-year college degree\4-year college degree or Postgraduate degree

7. What state do you live in?
   Dropdown with states – listed alphabetically; including District of Columbia and "None of the above"
8. In what year were you born?
   1947 or earlier\1948–1967\1968–1982\1983–1994
9. What is your race or ethnic group?
   White\Black\Hispanic\Other