

# Measuring Public Opinion via Digital Footprints

Roberto Cerina  
`roberto.cerina@nuffield.ox.ac.uk`  
Raymond Duch  
`raymond.duch@nuffield.ox.ac.uk`  
Nuffield College, University of Oxford

*Nuffield College, 1 New Road; Oxford; United Kingdom.*

---

## Abstract

Do digital traces accurately reflect individual preferences? Can signals from social media be used to measure public opinion? This paper provides evidence in favour of these hypotheses. We test a regression and post-stratification strategy that combines samples of digital traces with a stratification frame containing individual-level socio-economic data, in order to generate area forecasts of the outcome social phenomena of interest. In our example, we forecast the two-party vote of Democrats and Republicans in the 2018 Texas congressional district and Senate election. Our implementation assumes we can observe, and sample, individuals signaling their preference by favoring one virtual location over another. In our case, visiting Democrat versus Republican Facebook pages during the election campaign. Over the course of seven weeks preceding the mid-term elections we generate vote share forecasts which do not use any traditional survey data as input. Our results indicate that individuals leave digital traces that reflect their preferences.

*Keywords:* Social Media, Opinion Polling, Matching, Stratification, Voting

---

## 1. Introduction

Online social media increasingly dominates political discourse. An increasingly large percentage of the population leaves distinct digital traces that signal their political preferences. A challenge is mapping these digital traces to the general population and predicting behavior. We propose one method for accomplishing this goal and implement a “proof of concept” demonstrating that observing and recording digital traces can be the basis for predicting political behavior in the population. With recorded digital traffic to candidate Facebook pages, we forecast the 2018 mid-term elections in the state of Texas. Election forecasts are an ideal vehicle for assessing whether digital traces are a useful tool for measuring public opinion because the outcome of interest, the party or candidate vote share, is observed after the election. We assess the accuracy of digital trace forecasts by comparing actual outcomes in the 36 Texas congressional races (and one Senate race) with our predictions. Texas was selected for this “proof of concept” because it has a large number of diverse districts of varying competitiveness.

The post stratification dataset for generating the forecasts is an enumeration of all registered voters in the state of Texas. Our innovation is to construct a “virtual” sample consisting of voter preferences unobtrusively measured by observing who visits Republican and Democratic candidate Facebook pages. Using basic socio-demographic characteristics of these Facebook users, we map their revealed party preferences to the cells in the post stratification dataset. We then implement a random forest to model preferences and generate predicted vote share for the parties in each congressional district. The essay provides important insights into how digital trace can inform models predicting human behavior. It also highlights the challenges associated with collecting these digital trace data.

There are two important features of the prediction and post-stratification approach we implement. First, our post stratification dataset contains individual records for a significant proportion of the population. These are publicly available voter registration and vote history records. Secondly, our sampling frame is unique in that it is a virtual sample of

digital traces – we populate the sample with individuals we observe interacting with either a Republican or Democrat Facebook page. Rather than measuring self-reported preferences, we measure revealed preference.

The cells in our post stratification dataset reflect the individual characteristics that we believe determine vote choice. By generating a sufficiently large sample of virtual partisan decisions we can, in principle, estimate the partisan composition of each of these population cells. We do this by matching our observed virtual decision makers to the cells in the population frame – so, for example, the digital trace of a young black Republican auto-worker could be matched to an appropriate cell in our state of Texas population frame. We use the R[1] package `fastLink`[2] to match our virtual sample of digital traces to the voter registration list, based on names, sex and city of residency; which resulted in a success rate of approximately 53% percent.

We demonstrate that “digitally” revealed preferences are strongly correlated with behavior. Most individuals observed visiting either a Republican or Democrat candidate’s Facebook page voted in either a Republican or Democratic primary. Our sample’s digital partisanship is strongly correlated with the partisan primary in which they voted, as well as their imputed partisanship status on their voter registration record.

The partisan composition of individual cells in our sampling frame is informative to the extent that the likelihood of interacting with political content on social media is independent of voting preferences within the voting groups. A comparison of the social media partisanship of our virtual sample cells to the recorded partisanship of our population cells suggests that our independence assumption is roughly accurate for partisan voters. This independence allows us to generate an informative distribution of Republican and Democratic partisans within each cell of the state post stratification dataset.

For each cell in the post stratification dataset we estimate a vote probability based exclusively on the digital trace of individuals visiting candidates’ Facebook pages. The estimation is a novel variation of Multilevel Regression with Post-stratification (MRP)

[3, 4, 5, 6]. We employ a Probability Machine [7] which is a random-forest based algorithm implemented via the `ranger` [8] package. We then weight these probabilities by the likelihood of individuals turning out and aggregate them to obtain estimates of district and state level support.

The pre-election vote share predictions for both Senate and congressional districts performed satisfactorily when benchmarked against actual election results. The digital trace forecast called races correctly for the Senate and all but 3 congressional districts (7, 32 and 23). Absolute error was high due to significant attenuation bias, something to be expected in MRP applications [6], but which was likely exacerbated by our sampling procedure, modeling choices and high levels of noise in the digital sample. We correctly-called twenty-nine of the thirty-two contested congressional races and one Senate race, coming close to matching the performance of `FiveThirtyEight` [9, 10]<sup>1</sup>, though again matched less well in terms of absolute error. Donald Trump’s 2016 vote share at the congressional district level gives a lower absolute error, as a predictor of two-party Republican vote share, than either the digital trace forecast or `FiveThirtyEight`.

These results suggest that a virtual sampling strategy is promising. Forecasts based on sampling revealed preferences from digital activity on social media perform similarly to those based on conventional sampling methods. The essay describes how to implement a digital trace forecast and proceeds as follows: We first describe how we generated a sample of partisan digital traces of eligible Texas voters. A subsequent section indicates how we matched the virtual sample to the Texas population frame. A third section describes the random-forest estimation of Republican and Democrat voting probabilities for eligible voters in all 36 Texas congressional districts and the Senate. This is followed by our week-by-week mid-term election forecasts which are benchmarked against conventional polling forecasts. We conclude with a discussion of the challenges associated with using digital

---

<sup>1</sup>We benchmark against `FiveThirtyEight`’s polls-only model, and compare our weekly estimates to their daily estimates averaged by week.

trace data to forecast population behavior.

## 2. The Virtual Sample and Post Stratification Data

Research designs explicitly incorporate prediction and post-stratification strategies with the aim of improving the consistency and precision of estimated outcomes for subsets of populations. The technique is frequently employed for estimating political preferences and outcomes for sub-units of national populations [11]. In fact there are quite diverse applications for MRP techniques [12].

The method has four fundamental components: 1) a post stratification dataset that defines the individual-level categories that predict the outcome of interest; 2) an individual-level model that includes these categorical variables and, possibly, aggregate-level variables that predict the outcome of interest; 3) a sampling frame that indicates how the individual-level observations will be collected; and 4) a post-stratification algorithm. The goal is to map an estimated quantity (such as the likelihood of voting Republican) to each of the cells in the population frame (high educated black women in congressional district 32, for example). Simply multiplying this estimated likelihood times the population for this cell gives us the estimated Republican and Democrat voters (for this cell).

Predicted outcomes for each cell are generated with coefficients estimated from individual-level models. Current practice is to estimate these individual-level coefficients with relatively large survey samples of the population of interest [11, 13]. Some evidence suggests that the performance of post-stratification estimation varies significantly depending on the features of these survey samples [4]. On the other hand, there is evidence that post-stratification performs relatively well with unconventional sampling frames. A case in point is the MRP forecast of the 2012 election outcomes based on an unconventional convenience Xbox user sample [3].<sup>2</sup> In fact, individual-level estimates from highly skewed sampling

---

<sup>2</sup>Wang et al. administered a daily opinion survey to a very large sample of Xbox users. Their sample was extremely skewed: it over-represented Romney partisans, the young, the highly educated, and third-

frames are the basis for many MRP studies [12, 14]. A key factor in the success of these MRP estimations is sample size: MRP estimation needs to be powered in order to accurately estimate vote likelihoods of sub-categories. A smaller sample, generated in the same manner, is more likely to violate the ignorability condition and hence less likely to correctly calibrate rare sub-categories of voters. Hence, unconventional sampling frames as the basis for MRP can work so long as the sample sizes are very large.

### 2.1. The Texas “Virtual” Sample

The MRP techniques invite the exploration of novel sampling strategies on which to base post-stratification estimation. In this spirit, we propose a non-probability quota sample that consist of social media subscribers. A “virtual” sample is entirely unobtrusive. Individuals are never asked to report their preferences or behavior. Rather their choices on social media are observed, unobtrusively. In our case, we observe Facebook users commenting on partisan Facebook pages.<sup>3</sup> In this particular digital sampling frame we restrict ourselves to Facebook pages – our simple requirement from these pages is that they attract political partisans and there is evidence to this effect [15].<sup>4</sup>

Our digital sampling strategy consisted in research assistants (RAs)<sup>5</sup> monitoring the Facebook pages of all Texas Republican and Democratic candidates for congress and the Senate – a detailed description of the system followed by the RAs is available in Algorithm 1 in the Appendix, and summarized as follows: a total of 68 candidate Facebook pages were monitored (see Table A1 in the appendix for details on these pages); for the eight-week

---

party voters. Despite this, their MRP estimates closely tracked the published polls, outperforming them in predictive power within the last few weeks of the campaign.

<sup>3</sup>The U.S. has about 140 million Facebook subscribers and Texas has about 17 million. In both cases these represent about 43 percent of their respective populations <https://www.internetworldstats.com/stats26.htm>.

<sup>4</sup>There is also evidence that sentiment expressed on candidate Facebook pages correlates with election outcomes although this is of less relevance to the MRP predictions we are generating in this forecasting exercise [16].

<sup>5</sup>Research assistants collected on average just over 325 traces per day, and were paid roughly \$12 per hour. The average cost per digital trace was \$0.4, for a total of just over \$6,000 over 48 days. Up to four different RAs were used on any given collection day.

period leading up to the election RAs were tasked on a daily basis to collect up to 30 digital traces from the relevant pages. The RAs would collect information on Facebook users who left a digital trace expressing support on a candidate’s page, where digital support was deemed to come from a like, love or explicitly positive comment under a post published on the candidate’s page. Whether a comment classified as ‘explicitly positive’ was left up to the RAs to decide based on their qualitative judgment. RAs were also asked to note, in a **Qualtrics** survey instrument, basic information regarding these Facebook users – in particular, Facebook ID, gender, and current address and home address if available. These Facebook users are categorized as partisans based on the pages they express support for and then are added to our virtual convenience sample.<sup>6</sup> The assumption here is that visiting a candidate’s Facebook page and expressing a positive sentiment is a strong signal of voting intention. Again, there are recent findings suggesting that Twitter sentiment is correlated with partisan voter registration data [17] and that Facebook likes predict reported vote choice [18]. Post-stratification estimation is based on this partisan virtual sample.

A virtual sample can consist of any digital trace that reflects a choice amongst different virtual destinations or options. It could be, for example, liking an Uber versus Lyft Facebook page or responding positively to competing Instagram campaigns. We argue that these novel sampling frames can be the basis for quite robust MRP-like estimations. The 2018 mid-term congressional and Senate elections in Texas provided an opportunity to evaluate this alternative MRP estimation strategy. Over the course of the 7 weeks preceding the November 6, 2018 mid-term election day we collected 15,683 digital traces. Our innovation is to map voting probabilities estimated from a “virtual” Texas sample to a well-defined Texas population frame.

---

<sup>6</sup>Our research assistants were instructed only to consider Facebook users of voting age, i.e., 18 years or older. It is of course possible that we collected social media profiles of individuals who were not eligible to vote – for example, non-citizens or individuals below the age of 18. Virtually all of these ineligible voters would be screened out of the sample once these Facebook profiles were matched to the Texas voter files. The possible very rare exception would someone who died or was convicted of a felony after the vote registration data were collected.

Our Texas population frame has 38,880 cells. For each of the 36 Texas congressional election districts we disaggregate the population by registration-file partisanship (three categories), gender (two categories), age (six categories), ethnicity (five categories) and education (six categories). Our initial step, though, is to match individuals who visit a particular partisan FB page to one of 1,080 cells in the overall Texas population frame (prior to disaggregation to the 36 congressional districts). Hence, the “partisanship” of each cell is determined by the frequency with which we observe individuals, with a particular set of characteristics, visiting a partisan FB page.

The degree to which sampled partisan user counts in each cell reflect true cell preferences depends on whether the partisan members of the cell are represented in our quotas at the same rate they are in the population. For this to be the case we would need to assume that the conditional propensity to interact with social media is independent of partisanship. This is an independence assumption that is unlikely to be strictly observed in practice. In order to directly evaluate this assumption we would need to know “true cell preferences” or the true conditional propensities to interact with social media. We know neither and determining the precise magnitude of these deviations is beyond the scope of this particular project.

Of course, most sampling frames, convenience or not, will violate, to some extent, this independence assumption. The best we can do here is benchmark our forecast to these other conventional forecasts. To the extent that our digital trace sampling strategy generates similar forecasts as those employing more conventional sampling strategies, it is reasonable to conclude that violations of the independence assumption by digital trace sampling is no worse than the violations occurring with more conventional samples.<sup>7</sup>

---

<sup>7</sup>Earlier versions of this essay experimented with more “shallow” versions of this independence assumption. In one simulation we assumed that the marginal distribution of the two-party votes was 50-50. We then proceeded to re-sample within each cell accordingly to ensure this marginal likelihood was preserved. By balancing the marginal voting likelihood, the within-cell distribution of votes is as-if independent of the propensity to interact with social media. On balance these variations on the independence assumption did not generate significantly different results and in some cases reduced the accuracy of the forecasts.



A second concern is that the size of the virtual quotas is large enough to capture relevant cell differences in partisanship. We tune our sample size to ensure this is true for large categories in the voting population, under a probability-sampling scenario (a limiting best-case for our application). To tune sample size we assume a sampling distribution for cell counts of partisan voters of the following form:

$$\mathbf{n} \sim \text{Multinomial}(p_{1,r=1}, \dots, p_{G,r=1}, p_{1,r=0}, \dots, p_{G,r=0}, N); \quad (1)$$

where  $\mathbf{n}$  represents the vector of sampled counts for voters belonging to cell  $g$ , where  $g = 1, \dots, G$ , and expressing a preference for party  $r$ , where  $r = \{0, 1\}$  (indicating support for the Republican party). This set up enables us to leverage *worst-case-scenario* sampling for a multinomial distribution following the recommendations of Thompson [19]. We need to specify the nature of our population frame; in our case, indicating the relevant socio-economic categories in our population frame that predict vote choice.

We set the sample size to ensure that we can have sample frame cells that represent at least 2.5 percent of the population<sup>8</sup>. In order to ensure a probability of at least 0.9 that all estimates of the multinational parameters are within 0.025 of the population proportions, we need a sample size of at least  $N = 1610$ . Note here that 2.5% refers to the whole Texas population, as our cells will be collected at the state-wide level. Although we make use of

---

<sup>8</sup>By setting the expected distance from the population parameter  $d = 0.025$  with a given probability  $(1 - \alpha)$  we are essentially saying that if a given group's true proportion is 0.01, there is a good chance that the sample estimate for this quantity will be zero; hence our sample is only powered to accurately estimate group proportions which are around 2.5% of the population. This rough equivalence between  $d$  and the size of the smallest groups we are powered to estimate results from the tiny size of the true population proportions; as the groups approach zero, we need  $d$  to also approach zero to accurately capture their population levels. Moreover, given we are sampling to eventually derive Republican and Democrat voting likelihoods within small groups (say those that make up 0.1% of the population), small sampling errors can lead to huge swings in within-group voting likelihoods, and ideally we'd want  $d$  to be smaller than the group size. Note that 2.5% is significantly higher than  $1/1080 \approx 0.1\%$  of the population, which would allow our sample to be theoretically tuned to accurately estimate every single voter sub-groups. Instead we opt to be powered for the largest voting groups, here defined as groups making up 2.5% or more of the Texas voting population, and assume the smaller groups will be randomly distributed between the two major parties.

sampling theory to inform a-priori expectations regarding sample-size for desired power, this is still a non-probability convenience sample due to the sampling strategy and nature of the digital locations.

Our sampling strategy is dynamic, designed to calibrate weekly changes in voter preferences. Accordingly, we generate a weekly “virtual” sample of the Texas electorate. *Weeks-to-election* is the relevant time unit. Our monitoring of the election begins 7 weeks before election-day, hence monitoring the period starting on the 18<sup>th</sup> of September and ending on the 6<sup>th</sup> of November. Assuming a worst case scenario where each multinomial parameter is time-independent (and hence we need completely new information on every time period to estimate these accurately), we sample  $N$  as above for every week within our monitoring frame, for a total of  $N = 11,270$  individuals. The effective weekly sample size is further inflated with respect to the Thompson number to account for sample loss as a result of poor matching with the voter registry. While the assumption that the multinomial parameters are time-independent is certainly too strong, it will roughly be true for *swing* groups of voters. Treating each week as an independent sample should ensure the minimum sample size necessary to represent swing voters with reasonable precision. And stable partisans should be estimated precisely with this sampling strategy.

Procedures were implemented to avoid bias from systematic day-of-the-week effects on clustering of preferences<sup>9</sup>. We spread the collection of partisan voters evenly over the days of the week<sup>10</sup>. Initially we sampled, daily, 5 digital traces for each of the 68 FB pages (the total number of congressional and senatorial candidates). This generated a sample of 2,380 traces per week. As the campaign progressed we increased the sample size in order to promote higher collection rates for swing voters from a subset of the pages. The figure above is however only “potential” given that Facebook pages are not active at the

---

<sup>9</sup>For instance it is plausible that Fridays are a particularly bad day for the incumbent President’s party, as the “Friday news dump” may lead to heightened coverage of scandals or critique of partisan policy.

<sup>10</sup>The supplementary material contains Routine 1 that describes the data collection steps in detail

same rate. If no post is published on a given day, no user will have the opportunity to interact with the page. This creates a situation similar to non-response in an opinion poll. If this behavior is systematically correlated with voter characteristics used to estimate cell voting likelihood it can severely bias the estimates. Moreover, inactivity selectively reduces sample size, and makes it less likely for us to capture rare voter-types which may be over-represented in inactive pages.<sup>11</sup>

It is possible the same individuals leave multiple digital traces per week. For the following reasons we opt to consider each of these “draws” as a separate individual: i) we are not interested in individual-level predictions, but category-level; ii) some sub-categories are rarely sampled on social media, so this avoids wasting information; iii) some sub-categories are “swinging” during the campaign, and hence the same individual may leave different digital traces; iv) and because of post-stratification weighting, repeated traces will only be a problem if they are a consistently large outlier for the given cell.

In total, we sampled 15,683 digital traces; 53.6% were Democratic and 46.4% Republican. Our conjecture is that observing an individual’s digital trace is a particularly robust measure of preference. It is an unobtrusive revealed preference. And while, as we have seen in this section, constructing a virtual sample of these choices is challenging we believe that in many cases the benefits outweigh the costs. In order to implement the MRP, a necessary and challenging step is to match the virtual sample to the population.

## *2.2. Matching Sample to the Post Stratification Data*

We propose a novel strategy for matching our virtual sample with specific population frame cells – a critical step in generating the post-stratification estimates. Our virtual sam-

---

<sup>11</sup>Most pages we sample from correspond to a congressional district. This was done to ensure the interacting users would be expressing their preference over the congressional vote, as opposed to the senatorial or gubernatorial ballot. We did not include a congressional district variable in our likelihood estimation. If any district-candidate page is systematically less active than that of their district rival, district co-variables would capture artificial variation due to heterogeneous page-activity rates, and activity could be confounded with voting preferences.

ple includes a name, geographic location, possibly some limited demographic information and of course their partisan digital trace – whether they were observed on a Democratic or Republican congressional candidate Facebook page.<sup>12</sup> In order to match these digital preferences to the Texas population frame we need richer socio-demographic information on the individuals.

In our case, we get these richer profiles by matching our social media sample to the state-wide Texas L2 voter registration file.<sup>13</sup> This file has individual-level data for over 13 million registered Texas residents, ranging from voting history to socio-economic and demographic characteristics. Matching to the L2 file allows us to identify potentially non-registered voters and to generate turnout weights based on historical voting patterns. These administrative records of voters will have missing and incorrect data. A recent Pew Research study assesses the completeness and accuracy of the various variables included in these commercial voter registration files.[20] While there are clearly some short-comings to these data, for the purposes of post stratification they represent one of the best resources available.

There are other strategies for obtaining this socio-demographic information for individuals in the digital trace sample. Ideally, these predictive characteristics should also be collected unobtrusively and inexpensively. But having observed the outcome variable unobtrusively, one option would be to conduct conventional, obtrusive, surveys to obtain this information.

We obtain three variables from the Facebook individual profiles that can match to the L2 voter history file. Name is one variable. We use the R[1] package `humaniformat`[21] to parse the Facebook names and obtain individual entries for *First Name*; *Last Name*; *Middle Name* and *Name Suffix*; this allows us to exploit the high level of name detail of

---

<sup>12</sup>This information was collected systematically on a daily basis by a research assistant who was provided with a data collection protocol.

<sup>13</sup>L2 is a major supplier of voter data; see <https://www.l2political.com> for details.

our voter list counterparts. Secondly, we match Facebook profiles to the *City of Residency* record in L2. The closest social media counterpart to this variable is the *Current City* entry. When this is missing we use *Home Town* as a proxy. The location and name variables are recorded with considerable error. The social media source data has spelling mistakes and inaccuracies. These are further exacerbated by entry mistakes by research assistants who manually collect these characteristics. Because of the noise in these data we do not attempt perfect-matching on these variables. The final variable we match on is *gender*, which is unambiguously dichotomous in both datasets.

With the R package `fastLink`[2] we match our virtual sample to the voter registration list, based on names, sex and city of residency. `fastLink` leverages a probabilistic linkage Bayesian Mixture model, where similarity between variables can be a function of the Jaro-Winkler (JW) string-distance. It assigns a probability of match to each row in the first dataset, with respect to each row in the second. The model imputes missing values from the posterior distribution which enables us to match records even if entries for a given variable are missing in one or both datasets (in which case the similarity score between the remaining match-variables will determine the success of the match). The package is flexible enough to allow for mixed perfect-imperfect matching strategies; for instance, we match perfectly on *Gender*, and via string-distance on *First name*, *Last Name*, *Middle Name*, *Name Suffix*, *City of Residency*. For computational reasons, we set a high string-distance match threshold; namely, two strings are considered matched if the JW similarity is larger or equal to 0.99 (where 1 is an exact match). This allows us to match while ignoring minor inaccuracies. The overall matching threshold is set to the default 0.85, i.e. if the posterior probability of a match is equal to, or above, 0.85, the function returns the row index of a plausible match on the voter registration file, otherwise it does not.

From our social media sample of 15,683 digital traces, we successfully match 8,278 traces from 4,475 users to an entry in the voter-registration – this is a success rate of 52.8%. The matching uncertainty can be quantified by varying the matching threshold

from 0.75 to 0.99; this changes the number of matches from 8,739 to 7,322, highlighting 1,417 borderline cases. The False Discovery Rate (FDR), which is the proportion of false matches having overall posterior matching probability higher than the given threshold, is well controlled at 1.2%, suggesting the resulting matches are of high quality<sup>14</sup>. Holding the FDR close to zero is important – given our social media sample is quite small, we cannot afford to have the matching introduce further noise.

This match rate of 53 percent is respectable considering the identification information for each of our FB records can be quite noisy. Pew [20] reports the rate at which their American Trends Panel is matched to five different commercial voter files. In this case the identification information for each record is much cleaner. They report match rates of between 50 and 79 percent. A concern here is that we cannot assume that the unmatched records are “missing at random”. As Pew [20] suggest in their analyses of commercial data files, the unmatched records in their case were systematically younger, more mobile, less politically engaged, and less partisan. Hence our inability to match higher proportions of our digital sample could introduce bias into our vote preference estimations.

To assess whether digital traces accurately reflect political preferences, we assess whether those digital traces matched to the L2 voter file have a digital partisanship that a) corresponds to the imputed partisanship available in the L2 voter file; and b) matches the partisanship of the primary in which these individuals voted (if they cast a primary ballot).<sup>15</sup> Of those subjects with Republican digital traces, 71 percent were identified Republicans in the L2 database; a similar 83 percent of subjects who have a Democratic digital trace are Democrats in the L2 files. 88 percent of Republican digital traces voted in a Republican

---

<sup>14</sup>We do not report the False Negative Rate (FNR) calculated by the `fastLink::summary` function because the FNR is calculated relative to all potential pair comparisons, which leads to numbers quickly approaching 100% as the datasets become more imbalanced and less overlapping. The alternative `fastLink::confusion` function was computationally not feasible because the sheer size of our potential matches number in the billions. The function would in principle provide an FNR value relative to the reference population, which in our case is the social media sample. A visual inspection of the matches, as well as the outcomes of our predictions based on the matched sample, suggest the matching was high quality.

<sup>15</sup>The L2 voter file includes an imputed partisanship that is estimated by the commercial providers.

primary and 91 percent of Democratic digital traces voted in Democratic primaries. There is a high correlation between digital traces and partisan measures from our commercial voter registration files.

In a similar exercise Pew [20] find that between 62 and 85 percent of their American Trends Panelists who self identify as Democrat are identified as Democrat in the five commercial files they evaluate. For the Republicans, between 55 and 71 percent of those self-identifying in the ATP were identified as Republicans in the commercial files. And this congruence in the partisanship of our digital trace data and registration records is consistent with other similar efforts to measure partisanship and ideology with Facebook and Twitter digital traces [22, 17]. Hence, we are very comfortable concluding that digital traces represent revealed partisan preferences.

The frequency with which individuals leave partisan digital traces over the course of the monitoring period is also an indicator of partisanship. Those leaving multiple, as opposed to a single, digital traces for a party are more likely to be matched to the same party in the voter registration file. Out of the 2,130 matched voters that had Republican digital traces, 574 (around 27%) liked/loved/expressed support for a Republican page more than once during the monitoring period. Of the 2,392 matched voters that had Democrat digital traces, 716 (around 30%) liked/loved/expressed support for a Democrat page more than once. Amongst digital Republican supporters, the probability of being a Republican on the L2 file is 6.5% higher amongst individuals who left multiple pro-GOP digital traces, compared to individuals who left only one; similarly, amongst digital Democrats it is 4.5% more likely for individuals who left multiple pro-Democratic digital traces to be a partisan Democrat on the L2 file. Multiple likes seem to provide a cleaner signal of support, suggesting the frequency of digital traces can be leveraged in future work to obtain cleaner signals of support<sup>16</sup>.

---

<sup>16</sup>It was rare for voters to like both Democratic and Republican pages: we found that, out of the 4,475 matched users, only 47 liked posts on both pages. These appear to be overwhelmingly Democrats (70%), although the sample is too small to make meaningful inference. In principle, these voters could have been

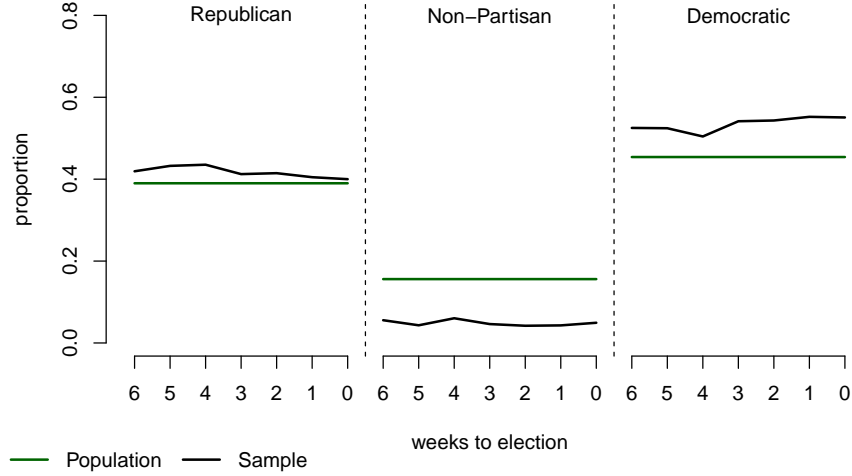


Figure 1: Population v. Sample comparison: partisanship.

Our final sample size is 8,278 digital traces. This is slightly below the Thompson number. Holding probability constant at 0.9 (that all estimates of the multinomial parameters are within 0.025 of the population proportions), we are powered (assuming equal sample size across weeks), at the weekly level, to estimate voter categories that are larger than 2.9 percent of the total population.

Figure 1 and Figures Appendix A.2 to Appendix A.5 in the Supplementary Appendix compare the socio-demographic profile of our matched sample to that of the population at large. With respect to *Partisanship*, we overestimate the percentage of Democratic and Republicans in the population and understate the percentage of Independents. This is an understandable result of sampling from partisan Facebook pages. We underestimate the young and overestimate the old and slightly over-represent women. Our virtual sample over-represents whites and slightly under-represents minorities. The education profile of the virtual sample is quite close to that of the population.

---

‘window shopping’ - visiting and liking multiple pages as part of their preference formation process. The number of voters who liked both sets of pages is quite small, suggesting that such a learning process is uncommon amongst users in our sample.



Just to recap, we matched a small convenience sample of individuals visiting partisan Facebook pages to the L2 database of Texas registered voters. We assess the extent to which convenience samples of this type are similar, on key variables, to the actual population of interest. Our conclusion is that reasonably calibrated virtual samples can be quite representative of the overall population.

### **3. Area Estimates of Partisan Support**

A critical step in the classic Multilevel Regression with Post-stratification (MRP) approach is to estimate a model of the outcome of interest based on a survey conducted with a sample of the relevant population [3, 4, 5, 6]. Coefficient estimates from these survey-based models facilitate the area-level estimations. In our case, the area estimates are of partisan support for each of the 36 congressional districts in Texas. We do not rely on survey responses to estimate our model of vote preference. Vote preference is an unobtrusively measured digital trace. And the explanatory variables in the vote choice model are all individual-level measures obtained from the matched L2 population file.

Our novel prediction and post-stratification estimation strategy for the 36 Texas congressional districts has three crucial steps: i) the identification of voter characteristics from the L2 file that would be used in the vote choice prediction model; ii) estimation, at the voter-category level, of vote choice conditional on the probability of vote turnout; iii) the weighting of these estimates by cell counts and summing over the area of interest to recover estimates of support.

#### *3.1. Variable Selection and Imputations*

Our MRP implementation relies on the L2 data base to provide all of the information required for constructing the classic population frame and also most of the information necessary for estimating the vote choice and turnout models. We rely on the virtual sample primarily to provide the measure of vote preference. Once this virtual sample is

matched to the L2 data base we can then use the rich set of L2 variables to predict their revealed vote preference.

From L2 we construct a core data set,  $\mathbf{X}$ . Variable selection for  $\mathbf{X}$  reflects the unique estimation strategy we propose here. We predict vote choice and turnout with a machine learning estimation strategy; specifically, we use random forests for both vote-choice and turnout. Hence, we impose no a priori functional form for the likelihood function for either models. We do, on the other hand, propose a slate of prospective variables that we anticipate are correlated with electoral behaviour. As Duch and Stevenson [23] point out, there is considerable agreement amongst voting behavior scholars as to what set of variables best predict electoral behaviour. These core variables include age, gender, imputed religion, household income, race, community size, partisan profiles (primary and past vote history, for example), and education. And these are predictors typically incorporated in U.S. vote choice models [24].

We include in  $\mathbf{X}$  additional variables from the rich L2 file – in total approximately 50 variables. Many of these  $\mathbf{X}$  will serve two other purposes in the estimation – impute missing values and predict voter turnout. Our search for these 50 candidates was motivated by the literature cited above – in particular, cells were built with the variables mentioned above, while strong correlates with these were included to motivate imputation and turnout models. Table Appendix B.1 in the appendix presents the full range of variables introduced in each model of electoral behaviour and in the imputation model. Note that our vote choice model is at the cell-level, while our turnout model is at the individual level. For the purpose of producing turnout estimates like in Figure 6, we simply sum the individual-level turnout probabilities across all the registered voters; the introduction of the roughly 50 individual level variables is justified by the desire to obtain better individual-level turnout propensities, and hence improve area-level estimates of turnout. In summary, the selection of these additional variables was guided by how well we anticipated they would predict turnout and perform in the imputation model.

For each  $h$  registered voter in data set  $\mathbf{X}$  we have  $\mathbf{x}_h$ . Missing values in  $\mathbf{X}$  are imputed with a random-forest multiple-imputation strategy implemented via the packages `ranger`[8] and `missForest`[25]. These provide a flexible non-parametric framework for imputing mixed-type data. Imputation error is estimated via calculating the Out of Bag (OOB) error for each imputed covariate, at each iteration. OOB error is roughly zero for all imputed variables suggesting that, if the assumption that data is *Missing at Random* holds, the observed data contains nearly all the information needed to complete individual records. This results in a completed dataset  $\mathbf{X}^I$  with voter-specific characteristics as  $\mathbf{x}_h^I$ . From  $\mathbf{X}^I$  we derive two datasets:  $\mathbf{Z}^I$ , indexed by  $h$  as above and  $\mathbf{X}^{I+}$ , indexed by expanded voter identifier  $h^+$ .

*Post Stratification Dataset.* All observations from the L2 voter registration file are included in the population frame. Hence,  $\mathbf{Z}^I$  is of the same length as  $\mathbf{X}^I$ . The covariates in  $\mathbf{Z}^I$  are modified. First, they are reduced in number. Secondly, the covariate values are recoded so they uniquely identify voter-categories of interest. This results in a feature space,  $m$ , that is simply the product of the number of covariates and covariate values. We define the voter category  $C_g$ , for categories  $g = 1, \dots, G$  as a unique realization of the set of variables which compose  $\mathbf{Z}^I$ , i.e.  $C_g = \{Z_1 = z_1, \dots, Z_m = z_m\}$ ; then for a voter  $h$  in the registry,  $h \in g$  if  $\mathbf{z}_h^I = C_g$ . Table Appendix B.1 in the appendix shows the voter characteristics chosen, along with their size in absolute numbers and as a proportion of the population. The feature space must be shrunk to  $m$  to ensure we have enough power to accurately represent the resulting categories' vote likelihood (recall from Thompson [19] we are powered to capture groups of voters which make up roughly 2.9% of the registered electorate). The resulting voting categories are then used to characterize the cells of our population frame.

The voter categories  $\mathbf{C}$  will not include a geographical identifier, such as congressional district number, as these will be systematically correlated with non-response. Nevertheless, the cells of our stratification frame will be defined by the interaction between categories  $\mathbf{C}$  and congressional districts; we will then allow the district-category counts in our population

frame (i.e. the number of individuals from category  $g$  which inhabit district  $D = d$  for  $d = 1, \dots, 36$ ) to weight our category-exclusive (non-district) likelihood and produce different district-category predictions. We note again that two districts with exactly the same demographic composition, under this set-up, will have the exact same predictions.

*Turnout data.* We do not have a digital trace measure of turnout likelihood. To estimate voter turnout probabilities we therefore rely on individual level historical voting behavior from the voter-registration file. To model the voting record of individuals over multiple years, the voter registration dataset is “expanded” such that each voter is “observed” over each of these years.  $\mathbf{X}^{\text{I}^+}$  is the expanded version of the completed voter sub-space, deployed for estimating turnout probabilities; it is constructed by using the L2 covariate “voted in year \_\_\_\_” as outcome variable, transforming  $\mathbf{X}^{\text{I}}$  into *long* format since, for any individual, we may observe their turnout in multiple elections.

### 3.2. Likelihood Prediction Models

The likelihoods predicting vote choice will employ these two datasets,  $\mathbf{Z}^{\text{I}}$  and  $\mathbf{X}^{\text{I}^+}$ , with complete imputed values. We will omit the imputation overscript “I” and simply refer to relevant completed datasets as  $\mathbf{X}$  and  $\mathbf{Z}$ . We estimate the joint probability of an individual in category  $g$  voting for the Republican candidate ( $R = 1$ ) and turning out on election-day ( $T = 1$ ) [6]:

$$P_g(R = 1, T = 1|C) = P_g(R = 1|T = 1, C) \times P_g(T = 1|C) \quad (2)$$

We again rely on the **ranger** package to implement a probability machine [7] composed of  $B^T$  trees. The terminal node of each tree will represent the relative frequency of respondents having the relevant voter-category characteristic. This is a random forest tuned to have probabilities in the terminal nodes of each of its trees, and standardized estimates producing an output-matrix with elements between zero and 1, and rows summing to 1. Random forests improves MRP estimation in two respects. First, the forest estimation

outputs the best<sup>17</sup> non-linear function of its inputs without us imposing a-priori functional specifications. If the input is not *important* [26] it will be consistently ignored by each decision tree. Second, from a practical stand-point, ranger-implemented forests are faster than almost any other machine available – given the size of our population frame, this was an important consideration.

### 3.3. Voter Turnout Estimation

Conditioning vote probabilities on the likelihood of turning out in an election is a critical element of area forecasts of election outcomes. Recent elections, such as the 2016 U.S. Presidential race, have highlighted the importance of estimating accurate turnout probabilities [13]. Efforts to model turnout probabilities typically rely on self-reported vote (either history or intention). Their advantage is that they capture the short-term temporal dynamics in turnout propensities. The Kiewiet et al [13] turnout model based on 2016 pre-election polls better predicted the demographic composition of voter turnout in the Presidential election than did models based on the 2012 Current Population Survey data. Given the importance of turnout composition for the 2016 Presidential election this explains in large part their ability to accurately predict electoral vote wins in all but one state.

Our turnout model relies exclusively on the L2 Texas voter registration data files. Voter registration files provided by state election authorities do not rely on reported turnout [27]. They avoid the well-known over-reporting of turnout in surveys [28, 29, 20] that can bias estimated turnout probabilities.<sup>18</sup> And there is evidence to this effect. Pew Research, for example, compared 2016 Presidential vote predictions weighted by turnout probabilities based on self-reports compared to commercial voter files. Predictions based on commercial

---

<sup>17</sup>According to minimization procedure of an out-of-bag prediction error

<sup>18</sup>Of particular concern is that this over-reporting reflects systematic sampling biases or variations in social desirability that are correlated with the demographic predictors that are included in models of turnout probability.

voter registration data outperform those that employ self-reported voting turnout weights [20]. These administrative files also have their drawbacks – examples include missing and erroneous data and the purging of the historical voting record of voters [30, 29]. Another disadvantage is that they are historical and hence do not help us estimate current, real-time, changes in the turnout propensities of key demographic categories.

We estimate, the probability of turnout,  $P_h(T = 1|\mathbf{x})$ , for all individuals in the complete voter file  $\mathbf{X}$ , where  $T$  is a random variable taking value 1 if the individual casts a ballot on election-day and 0 otherwise.  $P_h$ , is used as observation-weight in the estimation of the vote-choice likelihood, for each of our matched social media voters. Adjustments are taken to exclude voters who would have not been eligible to vote in earlier years; time-dependent characteristics are adjusted to reflect the year of the outcome variable; time-independent characteristics, such as race or sex, but also income or education, are assumed to stay constant over the years.

To estimate  $P_h$ , we train a random forest on the expanded turnout dataset  $\mathbf{X}^+$ ; the forest’s estimate of the probability of turning out on election day conditional on voter characteristics and being registered is defined as follows:

$$\hat{P}_{h+}(T = 1|\mathbf{x}^+) = \varphi^T(\mathbf{x}_{h+}^+) = \frac{1}{B^T} \sum_b^{B^T} \tau_b^T(\mathbf{x}_{h+}^+); \quad (3)$$

where  $\varphi^T$  represents the point estimate of a probability machine trained to predict turnout probabilities, and whose value is the average of  $B^T$  probability trees  $\tau_b^T$ . Having trained the model, we use it to output a prediction for the the probability of turning out for each member of the voting population:

$$\hat{P}_h(T = 1|\mathbf{x}) = \varphi^T(\mathbf{x}_h); \quad (4)$$

We extract the turnout probabilities of voters in our matched social media sample  $\hat{P}_s(T = 1|\mathbf{x})$  indexed by  $s = 1, \dots, S$  where  $s \subset h$ . This quantity will be used as the observation-weights

in our vote-choice model, effectively conditioning that distribution on turnout.

We calculate the empirical error distribution using the MSPE1 procedure from Lu [31]; this is designed to calculate the global mean-squared prediction error. We then assign a Normal distribution to the prediction error, and use the Root Mean Squared Predictive Error (RMSPE) as an estimator of the variance. Though the Normal distribution does not characterize the empirical distribution perfectly, it is useful to obtain reasonable prediction intervals; hence we describe the predictive distribution of turnout as follows:

$$P_h(T = 1|\mathbf{x}) \sim N\left(\varphi^T(\mathbf{x}_h^I), (\hat{\sigma}_{\text{RMSE1}}^T)^2\right). \quad (5)$$

The vote choice likelihoods in the area forecasts will be estimated for the  $g$  voter-categories. The probability of vote choice in these likelihoods will be conditional on the probability of turnout:  $P_g(T = 1|C)$ .  $P_g$  is the category-level probability of turning out on election day, which will be needed to calculate Equation 2. It is derived from  $P_h$ . We simply average across the voter categories identified by  $\mathbf{C}$  to obtain category-level estimates of turnout probabilities, as follows:

$$P_g(T = 1|C) = \frac{1}{\sum_{h \in g} 1(\mathbf{z}_h = C_g)} \sum_{h \in g \forall \mathbf{z}_h = C_g; \mathbf{z}_h \in \mathbf{x}_h} P_h(T = 1|\mathbf{x}); \quad (6)$$

where the outcome quantity is the average across a number of simulations from the predictive distribution of  $P_h(T = 1|\mathbf{x})$ .

### 3.4. Estimating Vote Probabilities

To estimate vote-choice probabilities we specify the outcome variable as the partisan digital trace  $R_s$  of voters  $s = 1, \dots, S$  where  $R_s = 1$  if the trace is Republican and 0 if Democrat. In total we have 8,278 individual traces from our digital sample that are matched to the voter registration data. With this matched sample, we train a probability machine as above to estimate the probability of voting Republican, conditional on the

individual turning out and the set of their voter characteristics:

$$\hat{P}_s(R = 1|T = 1, \mathbf{z}) = \varphi^R(\mathbf{z}_s|\hat{P}_s(T = 1|\mathbf{x})) \quad (7)$$

Here,  $\mathbf{Z}$  includes an identifier for each week to election  $W = w$  for  $w = 6, \dots, 0$ ; an identifier  $L = l$ , where  $l = \{1, 0\}$  indicating whether the trace at hand belongs to a state-wide or district level election; and the digital trace  $R = r$  where  $r = \{1, 0\}$ , which is our outcome variable. These new identifiers are also added to the categories  $\mathbf{C}$  used in the area forecasts. Note that the category-conditional turnout estimated previously is assumed to be constant across  $W$  and  $L$ .

Our next step in generating the area forecast is to assign a vote choice probability to each of the 1,080 cells in the Texas population frame. The trained forest is used to provide category-level predictions; for category  $g$  such that  $s \in g$  if  $\mathbf{z}_s = C_g$ :

$$\hat{P}_g(R = 1|T = 1, C) = \varphi^R(C_g). \quad (8)$$

Again, we estimate the global OOB Root Mean Squared Error MSPE1 procedure and encounter an empirical distribution which is approximated by a Normal probability distribution function:

$$P_g(R = 1|T = 1, C) \sim N(\varphi^R(C_g), (\hat{\sigma}_{\text{RMSE1}}^R)^2); \quad (9)$$

This section proposes a novel approach to the model estimation stage of classic MRP area estimation. First, the outcome variable, vote preference, is observed, unobtrusively, rather than self-reported. Second, these observed choices by individuals are matched directly to the population frame that includes an extensive set of explanatory variables, again not self-reported, that are employed in the prediction model. Third, the probabilities of vote choice for each of the population cells is estimated with a random forest machine that imposes no a-priori functional form on the model specification. This generates the two quantities needed to estimate the cell-level joint probability of voting Republican and



turning out, as specified in Equation 2.

### 3.5. Aggregation: Vote and Seat Predictions

The last step in the area estimation process is to generate congressional candidate vote predictions for each of the 36 congressional districts and vote predictions for the overall Senate race in the state of Texas. Accordingly, we have two kinds of aggregation to perform: district level and state-wide. The state-wide category level population counts are just the number of voters belonging to each category in the voter registration file:  $Q_g = \sum_h 1(z_h = C_g)$ . The state-wide area estimate for the vote share of the Republican party in the Senate election, with  $W$  weeks left in the campaign, is calculated as follows:

$$V_w^R = \frac{\sum_g P_g(R = 1, T = 1 | C, L = 1, W = w) \times Q_g}{\sum_g P_g(T = 1 | C, L = 1, W = w) \times Q_g}. \quad (10)$$

At the district level we count the number of voters in the intersection between a given category and a given district  $D = d$ ,  $\forall d = 1, \dots, 36$ :  $q_{gd} = \sum_h 1(z_h = C_g \cap D_h = d)$ . The district level estimates are the product of the following calculation:

$$v_{dw}^R = \frac{\sum_g P_g(R = 1, T = 1 | C, L = 0, W = w) \times q_{gd}}{\sum_g P_g(T = 1 | C, L = 0, W = w) \times q_{gd}}. \quad (11)$$

For each of the 36 Texas congressional districts this gives us a predicted vote (and vote share) for the Democratic and Republican candidates.

## 4. Results: Benchmarking Estimates

Figure 2 presents the Digital Vote Senate and congressional predictions for the election week period. Thirty-two of the 36 congressional districts fielded candidates from both major parties; 28 of these were seen as not particularly competitive prior to the election<sup>19</sup>

---

<sup>19</sup>Some races, such as district 24 and 31, ended up being more competitive than anticipated, as can be seen by the large error that `FiveThirtyEight` collected on these.

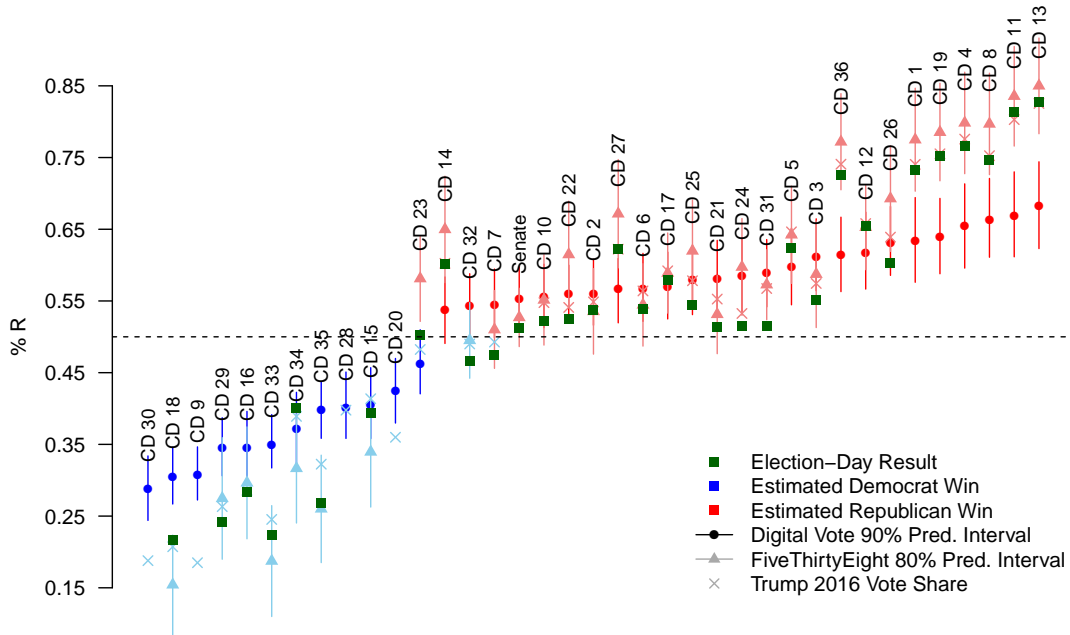


Figure 2: Election-week estimates of support for the Republicans, by area of interest; a comparison with *FiveThirtyEight.com* and actual election results is provided.

– and in these cases our calls match the actual election outcome. There were four a-priori competitive districts: Districts 7, 32, 23 and 21; District 7, and elected Democrats while we predicted Republican wins; District 23 elected (by a very small margin) a Republican whilst we predicted a Democratic win. We called District 21 correctly for the Republicans. The senate was also correctly predicted by our method. Of the 32 contested congressional races, we called 29 correctly. We also called the Senate race correctly. On balance this suggests our method is able to predict pluralities in most circumstances.

Of particular interest is how the Digital Vote weekly predictions for each of the Texas congressional districts compared to traditional survey estimates of party vote share. *FiveThirtyEight.com* aggregates public opinion surveys being conducted during this period and publishes daily forecasts of the vote shares for the Republican and Democratic congressional and Senate candidate contests in Texas. The *FiveThirtyEight* “light” model is based on a pro-

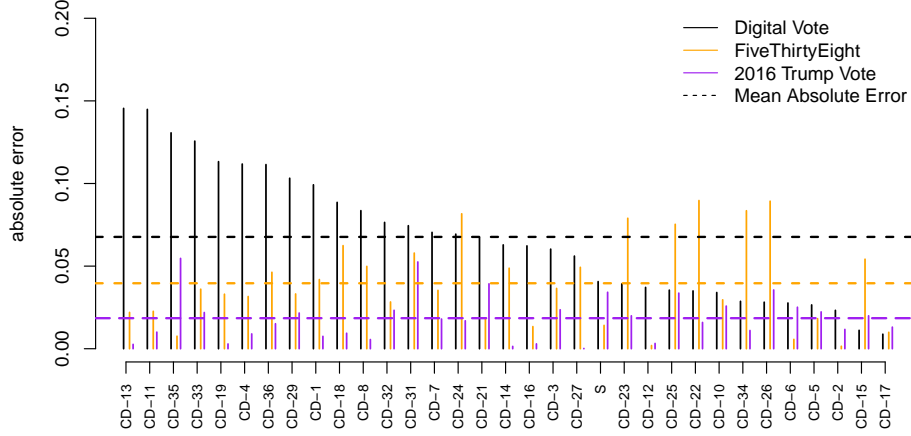


Figure 3: Mean Absolute Error: FiveThirtyEight, Trump 2016 and Digital Vote.

prietary aggregation of the latest public opinion polls. We focus on the “light” version of the site’s forecasts because it has no other information beyond polls and hence provides a more direct comparison to our method that emulates opinion polls using a digital sample.

The election week comparisons in Figure 2 suggest that the FiveThirtyEight aggregated results perform similarly to ours in the most competitive races: they call incorrectly call Districts 7 and 23, but correctly call Districts 32 and 21, as well as the Senate race. In terms of headline number of correct calls, they perform better than us by a single seat, with 30 contested congressional races and the Senate called correctly. Where our predictions fall short is in comparing Mean Absolute Errors (MAE).

Figure 3 summarizes the MAE for the two methods over the 36 congressional districts and the Senate seat. It also includes the MAE for a hypothetical forecast based on the 2016 two-party Trump share. These are based on the differences between the estimated two-party vote shares and actual vote share observed on election day.

There are two important messages in Figure 3. First, overall, prediction and post-stratification forecasts based on our virtual sampling strategy have high Mean Absolute

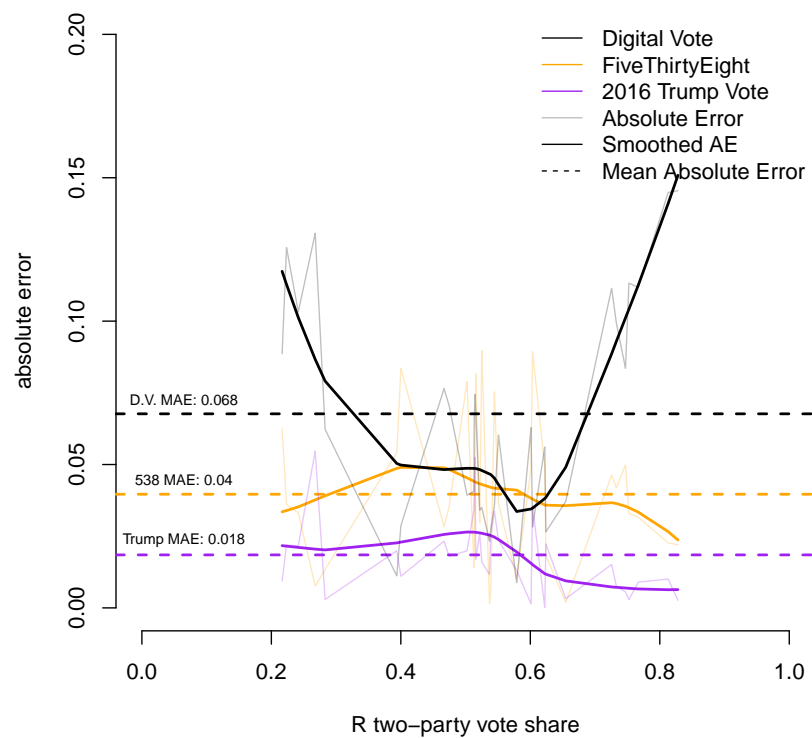


Figure 4: Mean Absolute Error, ordered by size of the Republican two-party vote share.

Error (MAE). Over all 32 districts where both major parties fielded candidates, the MAE is about 6.8 percentage points. The virtual sampling strategy does not perform as well as the FiveThirtyEight forecasts when we only consider Mean Absolute Error – their overall average is about 4 percentage points. FiveThirtyEight predictions are closer to the election-day vote share outcomes than those of Digital Vote. As mentioned earlier, the 2016 Trump vote obtains the lowest MAE.

Second, the Digital Vote suffers from significant attenuation bias. This is clear in Figure 4 where we plot the absolute error ordered by size of the Republican vote share. A rough U-shape emerges around the state-level predicted mean, clearly suggesting that absolute error is increasing where the voting population is less representative of the state average. Note though that the Digital Vote does perform significantly better around the state’s predicted mean, suggesting if we could address the attenuation bias, our performance would improve dramatically.

*Sub-category estimates.* Also of interest is the accuracy of the category-level predictions for our 1,080 sub-categories of voters. We compare the category-level predictions derived from our sample of 8,278 digital traces to those derived from regular surveys, holding modeling assumptions constant. We use two publicly available surveys from the *Texas Politics Project*<sup>20</sup> as benchmarks. In each poll 1,200 respondents were surveyed; the first poll was conducted between the 7<sup>th</sup> and the 15<sup>th</sup> of June; the second between the 15<sup>th</sup> and the 21<sup>st</sup> of October. These surveys are only relevant to the Senate race. We note that this survey sample, though it includes two separate polls, is heavily under-powered to accurately estimate the 1,080 cells in our model. Nevertheless, a comparison with our sample should show reasonable convergence on average, allowing for large sampling variability. To keep modeling assumptions as constant as possible, we use the reported voting history to condition vote-choice on turnout. The comparison is made between

---

<sup>20</sup>The Texas Politics Project’s polling data archive is available at: <https://texaspolitics.utexas.edu/polling-data-archive>.

estimates for the 2<sup>nd</sup> *week-to-election*, which is when the latest data is available for the surveys. The surveys included measures that had categories corresponding to those in Digital Vote, for Partisanship, Race, Age and Education. The exception is ‘Vocational Training’, which is an education category not surveyed by these polls. We can therefore produce comparisons for 900 voter sub-categories. Figure 5 compares estimate Republican votes derived by the digital trace to those observed in the Texas Project surveys.

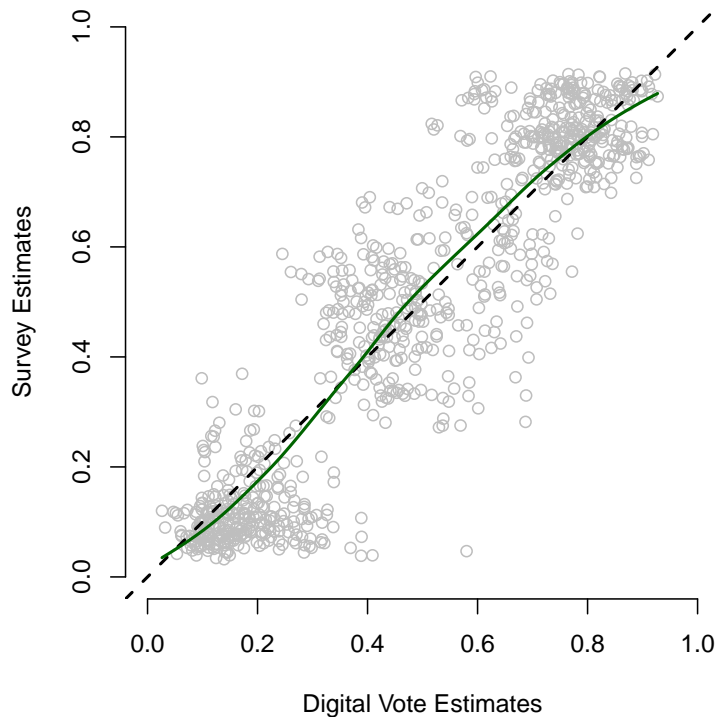


Figure 5: Voter sub-categories comparisons: Surveys v. Digital Vote.

Observations falling exactly on the 45 degree  $y = x$  line represent sub-categories for which the Texas Survey Project and Digital Vote predictions were identical. The plot indicates that Digital Vote matches closely the Texas Survey estimates; a slight over-estimation of the Republican vote seems to be concentrated around more Democratic constituency. This is consistent with the results of the area-estimates comparison in Figure 2. A LOESS

curve fit to the data closely tracks the  $y = x$  line suggesting that, at the sub-category level, digital traces reflect actual voting intentions.

*Turnout.* We estimate turnout probabilities from voting histories in the voter-registration file. Turnout as percent of registered voters in the Texas 2018 mid-term election was 53%; this represents an increase in excess of 19 percentage points over the previous 2014 mid-term election. As a result our turnout model that relied primarily on historical administrative data underestimated voting turnout. At the state-level we predicted 40.4% of registered voters to turnout while the actuals were 53.0%. Figure 6 benchmarks our district turnout forecasts against actual turnout statistics for the 36 districts and the overall state tally. The LOESS curve we fit to the data summarizes the strong correlation between our forecasted and actual turnout in the 36 districts. But for all districts actual turnout rates for 2018 were consistently higher than those predicted by our turnout model. Hence the LOESS curve lies above the 45 degree line. We also plot the actual 2018 district turnout percentages against turnout in the previous 2014 midterm election. Historical mid-term turnout significantly under-estimates the actuals for 2018. Consistent with the historically high 2018 mid-term turnout, we see in Figure 6 that turnout in the 2016 Presidential election is in fact a good predictor of 2018 turnout.

In our model, predicted vote choice, for any cell in the stratification frame, is conditioned on the likelihood of turning out to vote. While our turnout model performs poorly this is particularly problematic if there is systematic under-estimated turnout for particularly cells in the stratification frame (say, young female African Americans). We have some evidence from Texas Senate race exit polls suggesting this was not the case:<sup>21</sup> For every 100 self-identifying Republicans who cast their ballot in the Senate race, 83 Democrats turned out. In our digital sample, we predicted that 77 Democrats would show up to vote for every 100 Republicans who turned out. The digital sample had a small over-estimation of

---

<sup>21</sup><https://www.washingtonpost.com/graphics/2018/politics/voter-polls/texas.html>

Republican turnout; but far short of the 19 points turnout difference with 2014. Our conjecture is that the sharp rise in turnout was reasonably balanced across groups predisposed to vote Republican versus Democrat.

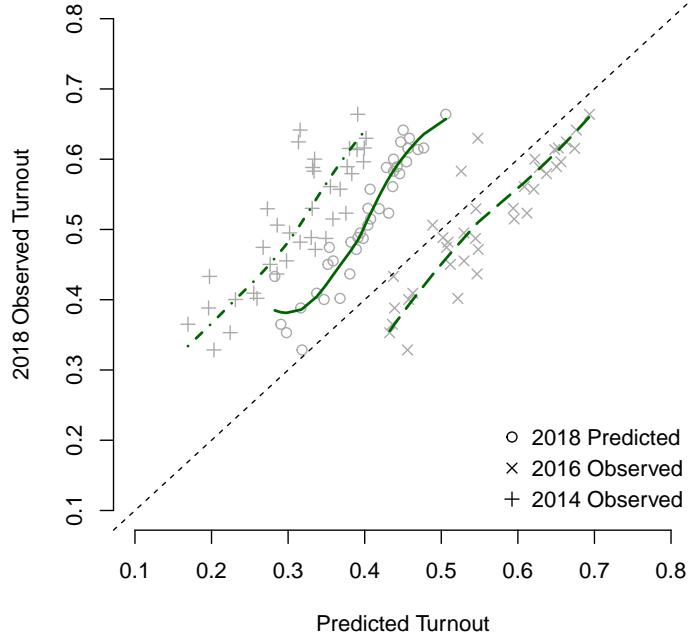


Figure 6: 2018 Turnout Predictions: Digital Vote versus 2018, 2016 and 2014 Actual Turnout.

*Costs and Benefits.* Generating our virtual sample of digital traces required monitoring visits to candidate Facebook pages and extracting relevant information from these pages. On balance the costs of generating virtual samples are comparable to those of conventional surveys. Of interest here is the cost of achieving comparable levels of predictive accuracy using our novel digital trace procedure relative to conventional survey methods. One metric is simply to ask what effective sample size would be required to generate the predictive accuracy of the Digital Vote and FiveThirtyEight estimates. Accordingly, we compare our sample size with that of hypothetical probability samples calibrated to achieve the same



level of accuracy. We generate these hypothetical comparisons by taking the observed point estimate from either Digital Vote or FiveThirtyEight and treating this as an informed prior for the purposes of calculating the sample size.

The sample size calculation for a proportion, in our case the share of the two-party vote going to the Republicans, is  $n = \left( \frac{z_{\alpha/2} \sqrt{\pi(1-\pi)}}{d} \right)^2$  with  $\pi$  set to a point estimate,  $\alpha$  set to 0.05 and  $d$  representing the desired margin of error. For each of the competitive races we forecast (where both parties were fielding candidates) we set  $d$  to equal the actual absolute errors from Figure 3, and calculate the sample size that independent probability surveys would have needed to achieve the same level of accuracy we did in each race. We generate these hypothetical sample sizes using the errors from both the Digital Vote and FiveThirtyEight. Sample size increases exponentially as the desired margin of error decreases – for example, the sample size required for a margin of error of 0.005 is 40,000. In some instances both our digital forecast and FiveThirtyEight have errors smaller than this implying sample sizes that could exceed half-a-million units. Hence, we cap the true effective sample size at  $n = 9600$ , or namely the effective sample size for a poll with a desired margin of error of  $d = 0.01$  under the worst-case-scenario  $\pi = 0.5$ . This lower-bound value for  $d$  is chosen to represent the maximum level of a-priori desired accuracy. Table Appendix C.1 in the Supplemental Appendix presents the hypothetical sample sizes for each district. Summing over the sample size needed for each district to achieve the observed level of accuracy, we find that independent probability surveys would have had to sample over 70,000 individuals to achieve FiveThirtyEight’s level of accuracy, and around 30,000 individuals to achieve ours. FiveThirtyEight’s level of accuracy is just under 2.5 times that of the Digital Vote, in terms of hypothetical sample size.

A conservative metric for assessing relative costs are those polls specific to Texas that FiveThirtyEight included in their model. These Texas polls conducted during the monitoring period would most strongly inform the FiveThirtyEight Texas congressional and Senate seat forecast [32]. These consisted of a total of 22 house surveys of over 11,000

respondents; 26 Senate surveys of over 25,000 individuals.<sup>22</sup> FiveThirtyEight’s aggregated estimates are less than 2 and a half times more accurate than our digital forecasts, for the Texas midterm district and Senate elections (the 30,000 versus 70,000 level of accuracy from Table Appendix C.1). If we just consider FiveThirtyEight’s house surveys then Digital Vote achieves this comparable level of precision with 4,475 unique Facebook users which is about 40 percent of the respondents in these 22 house surveys.

A potential benefit of these digital forecasts is that they provide added information to conventional forecasting strategies. An election-week regression of the results for the 32 contested congressional races and the Senate on the FiveThirtyEight polls-based predictions and our own forecast reveals a small, insignificant and negative coefficient on the Digital Vote. We conduct a similar regression estimation substituting simply the Trump 2016 vote for the FiveThirtyEight polls-based predictions – we do not add additional information to the Trump-based forecast (see Table Appendix D.1 in the Appendix). We also simply averaged our forecast with other available predictions on the assumption that forecasts may capture different information and hence will contribute to reducing errors; as a result, predictive power would be increased [33]. In fact, averaging either the Trump vote or the FiveThirtyEight forecast with our own does not make the forecast significantly better.

This forecasting exercise provides three insights into the costs and benefits of forecasting election outcomes with digital trace data. Digital trace data generate reasonably accurate predictions of congressional district vote shares. Secondly, the effective digital trace sample sizes required to attain these levels of precision are comparable to those of conventional surveys. Thirdly, the incremental information content of these digital predictions is limited. An interesting question is whether changes in design and estimation might enhance the predictive value added of digital trace approaches.

---

<sup>22</sup>These figures refer to the number of surveys conducted during our monitoring period.

## 5. Digital Challenges

This project explore strategies for incorporating unobtrusively measured digital traces into models predicting behavior. Our proof of concept is the vote for congressional and senate candidates in the 2018 mid-term elections in Texas. This is an ideal application because the accuracy of predictions can be calibrated against post-election outcomes. Digital traces predict pluralities at the district level; and individual level digital traces are highly correlated with actual voting behavior. On the other hand, our digital trace predictions incorporate little additional information above and beyond conventional forecast methods.

There are methodological challenges associated with collecting and modeling digital trace data. We briefly reflect on what can be done to enhance the precision of digital trace forecasting.

The adoption of a random forest estimation approach contributes to the precision of our forecasts. It allows us to minimize the number of parameters used to fit the model - each tree does not split at a given covariate unless the split reduces the out-of-bag mean squared error. At the same time though the tree forecasts are averaged to produce the forest prediction - effectively applying shrinkage toward the overall mean. MRP models though typically shrink coefficients towards the overall mean which tends to produce attenuation [6]. On the other hand, random forests are capable of identifying and modeling non-linear multi-way interactions [34], suggesting that our estimation strategy does capture relevant deep voter category effect. Hence, the estimation algorithm itself is not contributing to attenuation bias.

Attenuation results from omitting district-specific effects in our estimation procedure, hence precluding a-priori the identification of local heterogeneity. A solution to the attenuation bias is to include more information about the areas being forecast in the estimation; this is something that is regularly done in MRP applications to conventional surveys, and it has been shown to improve area estimates [35, 36, 37, 38].

We can improve the random forest estimation by including in the model complex inter-

action effects between area and demographics. For example, we would want to understand how the correlation between gender and digital vote preference differed in rural Republican-dominated districts compared to urban Democrat-dominated districts. This would require sampling larger amounts of area-specific digital trace data than we do in this exercise. As we pointed out earlier, our individual-level model of digital vote preferences is only sufficiently powered at the overall state level – hence, for example, for the gender effect we do not differentiate between rural Republican and urban Democrat districts. Our sampling based exclusively on official candidate pages limits the volume of digital activity we can sample. A larger overall sample-size of district-specific digital traces would have allowed us to capture district effects.

While it may not be feasible to collect sufficient digital traces for a specific district, it might be feasible to group districts into very homogeneous categories – lets say very noncompetitive rural Republican districts in the U.S. Assuming access to commercial voter registration files for the entire United States, one could then estimate vote preference equations for these “typical” U.S. congressional districts.

This would entail monitoring political preference expressed on a broader range of “API-friendly” social media (for example, Twitter) and matching the digital traces to a voter registration file so we learn, amongst other things, their congressional district. This strategy will generate limited information for some districts. They could be pooled with “high” information districts which are most ‘similar’ (in terms of demographic, geographic, political and cultural attributes) to the district we are interested in. We can consider these voters exchangeable conditional on some district-similarity parameter, allowing us to obtain exponentially more information at the district level.<sup>23</sup> These improvements alone, subject to resources and the availability of social-media data, would significantly reduce attenuation bias.

---

<sup>23</sup>A similar strategy was used by **FiveThirtyEight** to allow for information obtained via traditional opinion polling to flow from frequently polled districts to those less frequently polled.

By improving the quality of the digital trace sample we can reduce absolute error and the attenuation bias of our forecasts. There are four considerations here. One aim is simply increasing the size of the digital trace sample – including more individuals in the sample. Executing this in a cost-effective, and ideally automated, fashion is challenging.

A second factor that significantly contributes to the quality of the digital trace sample is complete and accurate information on the individuals in the sample. We were fortunate to have the complete voter registration data base for Texas and hence only needed basic information from Facebook accounts in order to match digital traces to the population frame along with their complete demographic profiles.<sup>24</sup> In most cases (certainly outside of the U.S.), researchers will not have this information. A challenge then will be obtaining a sufficiently complete demographic profile of the digital traces in order to match them to the population frame and estimate a reasonably well-specified prediction model.

A third consideration is the precision by which digital traces are matched to demographic profiles (commercial voter registrations, for example) and the population frame. The effective sample size, and its quality, increases with more precise matches. The matching algorithm contributes enormously here but also access to diverse population files can increase the matching of digital traces to demographic profiles.<sup>25</sup> The adoption of a two-stage matching procedure, one to ‘auxiliary files’ to clean the digital trace as much as possible, and finally to the L2 files, should be explored in further work.

Finally, as was alluded to earlier, we can improve the quality of the digital trace measure itself by measuring an individual’s partisan digital trace on multiple social media platforms.

---

<sup>24</sup>As Pew [20] points out though there is considerable variance in the completeness of these commercial files.

<sup>25</sup>Just to illustrate, in the Texas case we were unable to match a large number of digital traces to the L2 data file. We might have been able to match some of these to the Texas Appraisal District Files, assuming these had information which was consistent with L2, but slightly different so that it would have been closer to the Facebook data. For example, Appraisal District files might have had a slightly different permutation of name spelling and address specificity. With the additional demographic information from the Appraisal District Files it would ultimately become easier to match to the L2 file and therefore increase the overall digital trace sample size.

This would require observing the digital trace left by the same individual on different social media such as Facebook, Twitter, or WhatsApp but also multiple partisan digital traces by the same individual on a particular social medium over time. An important challenge here would be matching individuals to their various publicly available social media accounts. But as our analyses demonstrate this could be useful because the volume of an individual’s digital traces seems to correlate with strength of partisanship.<sup>26</sup>

*Conclusion.* We propose a novel MRP estimation strategy that combines samples of these digital traces with a population frame that has extensive individual-level socio-economic data in order to generate area forecasts of the outcome variable of interest. Our implementation assumes we can observe, and sample, individuals signaling their preference by favoring one virtual location over another. The digital trace in our case is visiting a Democrat versus Republican Facebook page during the election campaign. We face a challenge in measuring the socio-demographic characteristics of the individuals leaving these digital traces. We implement one strategy – matching individuals to quite exhaustive voter registration files – but there are a number of alternatives here. We demonstrate that even a relatively small virtual sample can be quite representative of the overall population. Finally, we train a random forest machine to estimate the probability of voting Republican, conditional on individual-level data from the complete voting history and registration data for Texas. Over the course of seven weeks preceding the mid-term elections we generate vote share forecasts for all 36 congressional seat contests (including “potential” votes if either party does not field a candidate) and for the Senate race. The forecasts do not use any survey results as input. The predicted vote pluralities are comparable to other conventional methods. Primarily because of attenuation bias, the absolute errors of our forecasts are greater than those of other methods. By optimising sampling to be theo-

---

<sup>26</sup>Although this should not detract from the important advantage of the Facebook medium because Facebook has in effect standardized the support procedure – a like or love is a much clearer signal of support than a retweet, or a friendship.

retically representative at the district level and by incorporating district-specific variables in our modeling, we could significantly reduce attenuation bias and significantly improve forecast precision.

## Appendix A. Sampling

---

**Algorithm 1** A description of the relevant steps in our sampling routine. The page space  $\rho$  included 68 partisan pages;  $N(M)$  was initially set to 5 for all pages, and later increased to 30 for the subset of pages we believed to be most informative of “swing” voters; relevant collected characteristics included *Digital Partisanship*; *Facebook Name*; *Current City*; *Home Town*; *Gender*.  $M$  was defined as any random pick of the  $N(M)$  *likes*, *loves* or *explicitly positive comments* on the relevant page, for the given day.

---

```
1: procedure GETUSERINFO
2:   let  $\rho$  a vector of relevant social media pages
3:   let  $N$  be a counting operator
4:   let  $M(\rho_{O_d})$  be the subset of users being collected per page, according to daily order  $O_d$ 
5:   let collect be a function of the input user, with output  $z$ , the set of relevant characteristics
6:
7:   # FOR EACH WEEK LEFT IN THE CAMPAIGN
8:   for  $w$  in  $W, \dots, 0$  do
9:
10:    # FOR EACH DAY WITHIN THE WEEK
11:    for  $d$  in  $7, \dots, 1$  do
12:
13:      # SAMPLE AT RANDOM A COLLECTION ORDER
14:       $O_d = \text{sample}(\text{from} = 1, \dots, N(\rho), \text{size} = N(\rho), \text{replacement} = \text{FALSE})$ 
15:
16:      # FOR EACH PAGE IN THE PAGE SPACE
17:      for  $O_d$  in  $1, \dots, N(\rho)$  do
18:
19:        # VISIT THE LATEST DAILY POST ON THE PAGE
20:        goto  $\rho_{O_d}$ 
21:
22:        # FOR EACH EXPLICITLY PARTISAN USER IN  $M(\rho_{O_d})$ 
23:        for  $i$  in  $1, \dots, N(M(\rho_{O_d}))$  do
24:
25:          # COLLECT INFORMATION FROM THEIR PUBLIC ABOUT PAGE
26:           $z_i = \text{collect } M_i(\rho_{O_d})$ 
27:
28:        end for
29:      end for
30:    end for
31:
32:  end for
33:
34: end procedure
```

---



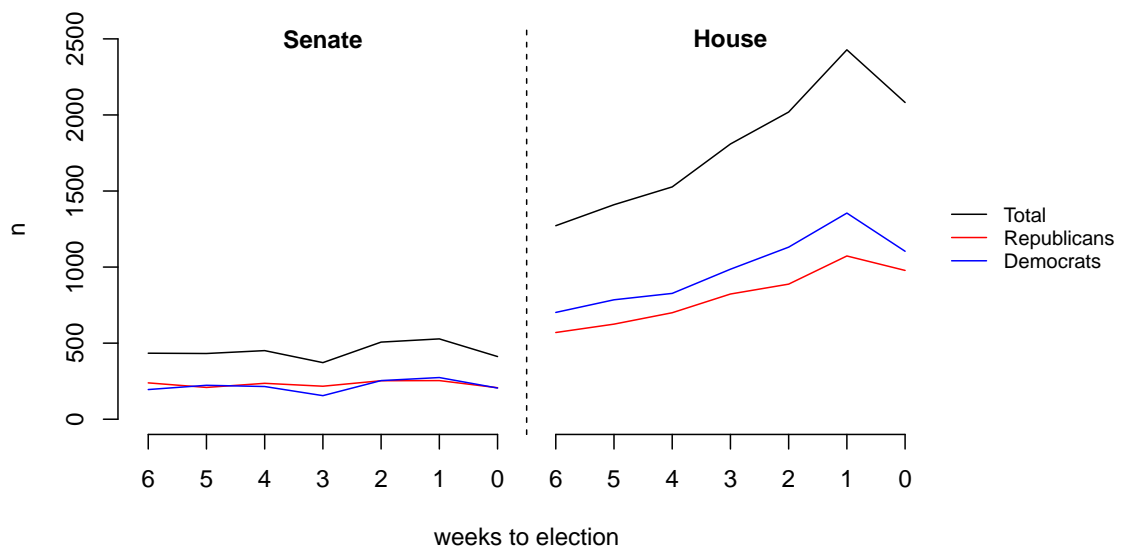


Figure Appendix A.1: Sample size of digital traces collected by party and election-type, over the monitoring period. The increasing trend in collection for the house is due to increasing collection for “swing” districts, in the hope of better filling competitive cells. Weaker numbers for “election week” are due to lower number of days available for collection

Table Appendix A.1: Sample summary by page.

Page Address	Party	Election Type	District ID	n	% Total
<a href="https://www.facebook.com/betoorourke/">https://www.facebook.com/betoorourke/</a>	Dem	US Senate	0	1521	9.70
<a href="https://www.facebook.com/tedcruzpage/">https://www.facebook.com/tedcruzpage/</a>	Rep	US Senate	0	1615	10.30
<a href="https://www.facebook.com/vote4McKellar">https://www.facebook.com/vote4McKellar</a>	Dem	US House	1	19	0.12
<a href="https://www.facebook.com/RepLouieGohmert/">https://www.facebook.com/RepLouieGohmert/</a>	Rep	US House	1	125	0.80
<a href="https://www.facebook.com/toddlittonforcongress/">https://www.facebook.com/toddlittonforcongress/</a>	Dem	US House	2	230	1.47
<a href="https://www.facebook.com/CrenshawforCongress/">https://www.facebook.com/CrenshawforCongress/</a>	Rep	US House	2	192	1.22
<a href="https://www.facebook.com/Lorieburchforcongress/">https://www.facebook.com/Lorieburchforcongress/</a>	Dem	US House	3	98	0.62
<a href="https://www.facebook.com/VanForTexas/">https://www.facebook.com/VanForTexas/</a>	Rep	US House	3	44	0.28
<a href="https://www.facebook.com/KrantzforCongress/">https://www.facebook.com/KrantzforCongress/</a>	Dem	US House	4	224	1.43
<a href="https://www.facebook.com/RepRatcliffe/">https://www.facebook.com/RepRatcliffe/</a>	Rep	US House	4	168	1.07
<a href="https://www.facebook.com/lancegoodenfortexas/">https://www.facebook.com/lancegoodenfortexas/</a>	Rep	US House	5	114	0.73
<a href="https://www.facebook.com/JanaLynneSanchezforUSCongress/">https://www.facebook.com/JanaLynneSanchezforUSCongress/</a>	Dem	US House	6	196	1.25
<a href="https://www.facebook.com/wright4congress/">https://www.facebook.com/wright4congress/</a>	Rep	US House	6	141	0.90
<a href="https://www.facebook.com/LizzieForCongress/">https://www.facebook.com/LizzieForCongress/</a>	Dem	US House	7	786	5.01
<a href="https://www.facebook.com/johnculberson/">https://www.facebook.com/johnculberson/</a>	Rep	US House	7	598	3.81
<a href="https://www.facebook.com/stevenforcongress/">https://www.facebook.com/stevenforcongress/</a>	Dem	US House	8	26	0.17
<a href="https://www.facebook.com/BradyforTexas/">https://www.facebook.com/BradyforTexas/</a>	Rep	US House	8	148	0.94
<a href="https://www.facebook.com/repalgreen/">https://www.facebook.com/repalgreen/</a>	Dem	US House	9	105	0.67
<a href="https://www.facebook.com/siegelfortexas/">https://www.facebook.com/siegelfortexas/</a>	Dem	US House	10	205	1.31
<a href="https://www.facebook.com/MichaelMcCaulTX/">https://www.facebook.com/MichaelMcCaulTX/</a>	Rep	US House	10	106	0.68
<a href="https://www.facebook.com/mike.conaway/">https://www.facebook.com/mike.conaway/</a>	Rep	US House	11	136	0.87
<a href="https://www.facebook.com/VanessaAdiaTX12/">https://www.facebook.com/VanessaAdiaTX12/</a>	Dem	US House	12	198	1.26
<a href="https://www.facebook.com/RepKayGranger/">https://www.facebook.com/RepKayGranger/</a>	Rep	US House	12	111	0.71
<a href="https://www.facebook.com/gregsagan2018/">https://www.facebook.com/gregsagan2018/</a>	Dem	US House	13	140	0.89
<a href="https://www.facebook.com/ThornberryForCongress/">https://www.facebook.com/ThornberryForCongress/</a>	Rep	US House	13	72	0.46
<a href="https://www.facebook.com/adrbell/">https://www.facebook.com/adrbell/</a>	Dem	US House	14	136	0.87
<a href="https://www.facebook.com/WeberForTexas/">https://www.facebook.com/WeberForTexas/</a>	Rep	US House	14	41	0.26
<a href="https://www.facebook.com/votevicente/">https://www.facebook.com/votevicente/</a>	Dem	US House	15	124	0.79
<a href="https://www.facebook.com/westley4Congress/">https://www.facebook.com/westley4Congress/</a>	Rep	US House	15	167	1.06
<a href="https://www.facebook.com/voteforveronica/">https://www.facebook.com/voteforveronica/</a>	Dem	US House	16	140	0.89
<a href="https://www.facebook.com/Seeberger1ForCongress/">https://www.facebook.com/Seeberger1ForCongress/</a>	Rep	US House	16	204	1.30
<a href="https://www.facebook.com/RickKennedyforCongress/">https://www.facebook.com/RickKennedyforCongress/</a>	Dem	US House	17	153	0.98
<a href="https://www.facebook.com/BillFloresForCongress/">https://www.facebook.com/BillFloresForCongress/</a>	Rep	US House	17	161	1.03
<a href="https://www.facebook.com/CongresswomanSheilaJacksonLee/">https://www.facebook.com/CongresswomanSheilaJacksonLee/</a>	Dem	US House	18	105	0.67
<a href="https://www.facebook.com/Ava-for-Congress-526965147465733/">https://www.facebook.com/Ava-for-Congress-526965147465733/</a>	Rep	US House	18	122	0.78
<a href="https://www.facebook.com/miguellevario19/">https://www.facebook.com/miguellevario19/</a>	Dem	US House	19	195	1.24
<a href="https://www.facebook.com/JodeyArrington/">https://www.facebook.com/JodeyArrington/</a>	Rep	US House	19	142	0.91
<a href="https://www.facebook.com/JoaquinCastroTX/">https://www.facebook.com/JoaquinCastroTX/</a>	Dem	US House	20	143	0.91
<a href="https://www.facebook.com/ChipRoyforCongress/">https://www.facebook.com/ChipRoyforCongress/</a>	Dem	US House	21	230	1.47
<a href="https://www.facebook.com/KopserforCongress/">https://www.facebook.com/KopserforCongress/</a>	Rep	US House	21	235	1.50
<a href="https://www.facebook.com/KulkarniforCongress/">https://www.facebook.com/KulkarniforCongress/</a>	Dem	US House	22	225	1.43
<a href="https://www.facebook.com/PeteOlsonTX/">https://www.facebook.com/PeteOlsonTX/</a>	Rep	US House	22	126	0.80
<a href="https://www.facebook.com/GinaOrtizJones/">https://www.facebook.com/GinaOrtizJones/</a>	Dem	US House	23	755	4.81
<a href="https://www.facebook.com/HurdForCongress/">https://www.facebook.com/HurdForCongress/</a>	Rep	US House	23	747	4.76
<a href="https://www.facebook.com/JanMcDowellDemocrat/">https://www.facebook.com/JanMcDowellDemocrat/</a>	Dem	US House	24	179	1.14
<a href="https://www.facebook.com/RepKennyMarchant/">https://www.facebook.com/RepKennyMarchant/</a>	Rep	US House	24	150	0.96
<a href="https://www.facebook.com/JulieForTexas/">https://www.facebook.com/JulieForTexas/</a>	Dem	US House	25	192	1.22
<a href="https://www.facebook.com/RepRogerWilliams/">https://www.facebook.com/RepRogerWilliams/</a>	Rep	US House	25	186	1.19
<a href="https://www.facebook.com/LinseyFaganTx/">https://www.facebook.com/LinseyFaganTx/</a>	Dem	US House	26	218	1.39
<a href="https://www.facebook.com/michaelcburgess/">https://www.facebook.com/michaelcburgess/</a>	Rep	US House	26	175	1.12
<a href="https://www.facebook.com/EricHolguin/">https://www.facebook.com/EricHolguin/</a>	Dem	US House	27	222	1.42
<a href="https://www.facebook.com/CloudforCongress/">https://www.facebook.com/CloudforCongress/</a>	Rep	US House	27	161	1.03
<a href="https://www.facebook.com/repcuellar/">https://www.facebook.com/repcuellar/</a>	Dem	US House	28	53	0.34
<a href="https://www.facebook.com/SylviaRGarcia/">https://www.facebook.com/SylviaRGarcia/</a>	Dem	US House	29	147	0.94
<a href="https://www.facebook.com/aronoffforcongress/">https://www.facebook.com/aronoffforcongress/</a>	Rep	US House	29	112	0.71
<a href="https://www.facebook.com/CongresswomanEBJtx30/">https://www.facebook.com/CongresswomanEBJtx30/</a>	Dem	US House	30	89	0.57
<a href="https://www.facebook.com/MJforTexas/">https://www.facebook.com/MJforTexas/</a>	Dem	US House	31	72	0.46
<a href="https://www.facebook.com/judgecarter/">https://www.facebook.com/judgecarter/</a>	Rep	US House	31	103	0.66
<a href="https://www.facebook.com/ColinAllredTX/">https://www.facebook.com/ColinAllredTX/</a>	Dem	US House	32	819	5.22
<a href="https://www.facebook.com/petesessions/">https://www.facebook.com/petesessions/</a>	Rep	US House	32	429	2.74
<a href="https://www.facebook.com/MarcVeasey/">https://www.facebook.com/MarcVeasey/</a>	Dem	US House	33	52	0.33
<a href="https://www.facebook.com/billups4congress/">https://www.facebook.com/billups4congress/</a>	Rep	US House	33	227	1.45
<a href="https://www.facebook.com/UsCongressmanFilemonVela/">https://www.facebook.com/UsCongressmanFilemonVela/</a>	Dem	US House	34	31	0.20
<a href="https://www.facebook.com/profile.php?id=100011088094658&amp;fref=mentions">https://www.facebook.com/profile.php?id=100011088094658&amp;fref=mentions</a>	Rep	US House	34	65	0.41
<a href="https://www.facebook.com/LloydDoggettTX/">https://www.facebook.com/LloydDoggettTX/</a>	Dem	US House	35	156	0.99
<a href="https://www.facebook.com/Davidsmallingforcongress/">https://www.facebook.com/Davidsmallingforcongress/</a>	Rep	US House	35	12	0.08
<a href="https://www.facebook.com/daynasteale36/">https://www.facebook.com/daynasteale36/</a>	Dem	US House	36	227	1.45
<a href="https://www.facebook.com/RepBrianBabin/">https://www.facebook.com/RepBrianBabin/</a>	Rep	US House	36	137	0.87

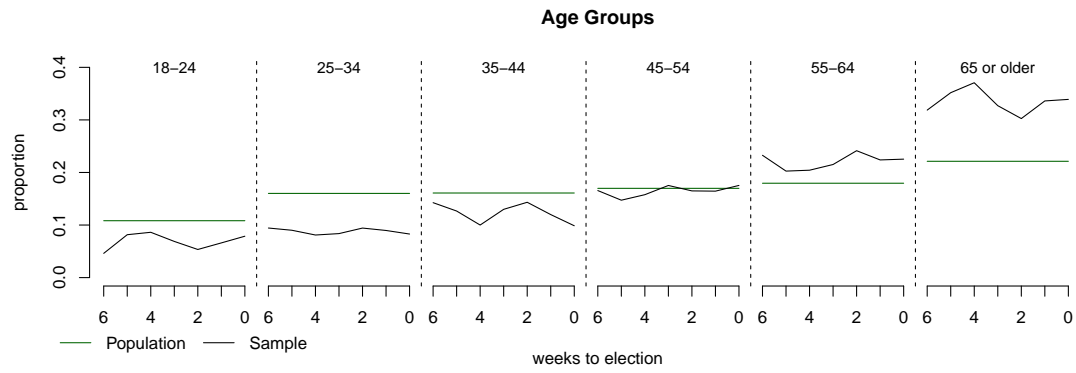


Figure Appendix A.2: Population v. Sample comparison: age.

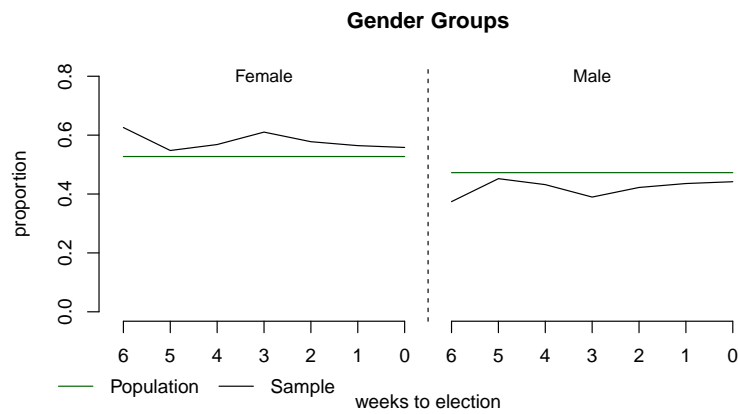


Figure Appendix A.3: Population v. Sample comparison: gender.

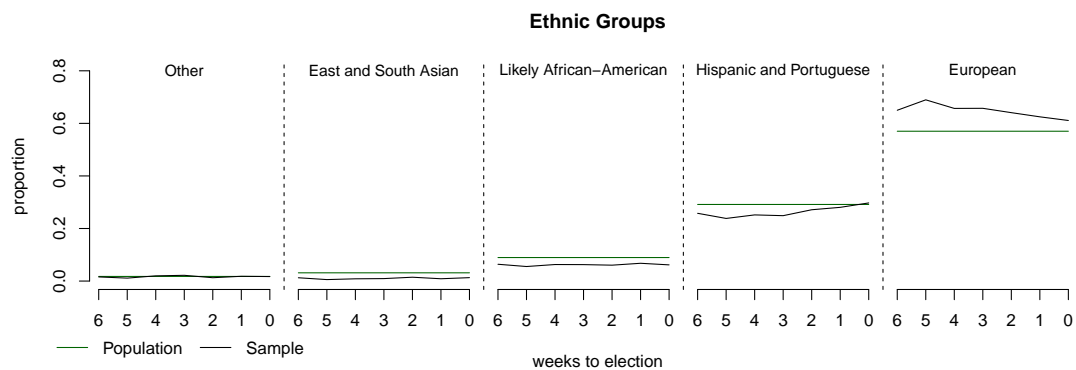


Figure Appendix A.4: Population v. Sample comparison: ethnicity.

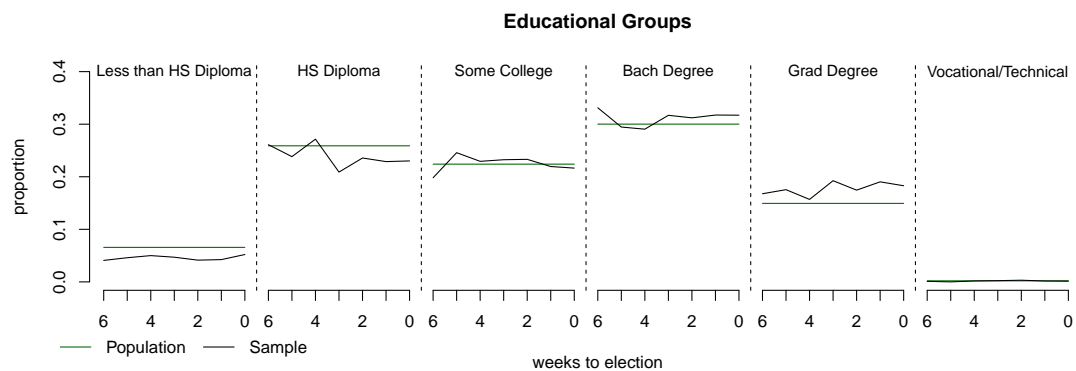


Figure Appendix A.5: Population v. Sample comparison: highest level of education.

## Appendix B. Variable Selection

Table Appendix B.1: Summary table of the reduced covariate-space  $\mathbf{X}$ .

Variable_Names	Class	Vote Choice Cell	$P(\mathbf{x} = \text{NA})$	Median/Mode
US_Congressional_District	factor		0.00	21
Voters_Gender	factor	✓	0.00	F
age_cat	factor	✓	0.00	65_or_older
EthnicGroups_EthnicGroup1Desc	factor	✓	0.06	European
CommercialData_Education	factor	✓	0.38	Bach Degree - Extremely Likely
income_cat_000	factor		0.03	(80,100]
Religions_Description	factor		0.51	Protestant
ElectionReturns_G16_Cnty_Margin_Trump_R	numeric		0.00	9
ElectionReturns_G12_Cnty_Margin_Obama_D	numeric		0.00	-11
ElectionReturns_G12PrecinctTurnoutAllRegisteredVoters	numeric		0.00	53
ElectionReturns_G14PrecinctTurnoutAllRegisteredVoters	numeric		0.00	32
ElectionReturns_G16PrecinctTurnoutAllRegisteredVoters	numeric		0.00	62
ElectionReturns_G12PrecinctTurnoutDemocrats	numeric		0.00	49
ElectionReturns_G14PrecinctTurnoutDemocrats	numeric		0.00	27
ElectionReturns_G16PrecinctTurnoutDemocrats	numeric		0.00	58
ElectionReturns_G12PrecinctTurnoutIndependentsAllOthers	numeric		0.00	24
ElectionReturns_G14PrecinctTurnoutIndependentsAllOthers	numeric		0.00	7
ElectionReturns_G16PrecinctTurnoutIndependentsAllOthers	numeric		0.00	34
ElectionReturns_G12PrecinctTurnoutRepublicans	numeric		0.00	69
ElectionReturns_G14PrecinctTurnoutRepublicans	numeric		0.00	48
ElectionReturns_G16PrecinctTurnoutRepublicans	numeric		0.00	76
General_2016_11_08_reg	factor		0.24	Y
age_cat_2016_11_08	factor		0.00	65_or_older
General_2014_11_04_reg	factor		0.22	N
age_cat_2014_11_04	factor		0.00	45-54
General_2012_11_06_reg	factor		0.39	Y
age_cat_2012_11_06	factor		0.00	45-54
CommercialData_BookBuyerInHome	numeric		0.78	2
CommercialData_AreaMedianEducationYears	numeric		0.07	12
CommercialData_EstHomeValue	numeric		0.04	175308
CommercialData_EstimatedAreaMedianHHIncome	numeric		0.07	71854
CommercialData_HHComposition	factor		0.76	1 adult Male & 1 adult Female
CommercialData_OccupationGroup	factor		0.46	Retired
Parties_Description	factor	✓	0.00	Democratic
VotingPerformanceEvenYearGeneral	numeric		0.10	60
VotingPerformanceEvenYearPrimary	numeric		0.16	0
VotingPerformanceMinorElection	numeric		0.08	0
Voted_in_Primary_2018_03_06	numeric		0.41	0
Voted_in_Primary_2016_03_01	numeric		0.42	0
Voted_in_Primary_2014_03_04	numeric		0.53	0
Voted_in_Primary_2012_05_29	numeric		0.57	0
Voted_in_Primary_2018_03_06_party	factor		0.80	R
Voted_in_Primary_2016_03_01_party	factor		0.71	R
Voted_in_Primary_2014_03_04_party	factor		0.87	R
Voted_in_Primary_2012_05_29_party	factor		0.87	R

## Appendix C. Results

District	Digital Vote Error	Digital Vote $n$	FiveThirtyEight Error	FiveThirtyEight $n$
Senate	4.06	575	1.42	4,782
CD-1	9.92	91	4.19	382
CD-2	2.32	1,752	0.15	9,555
CD-3	6.03	251	3.65	698
CD-4	11.18	69	3.16	618
CD-5	2.65	1,312	1.82	2,665
CD-6	2.77	1,230	0.57	9,527
CD-7	7.04	192	3.53	769
CD-8	8.36	123	4.99	250
CD-9				
CD-10	3.4	819	2.96	1,084
CD-11	14.49	41	2.27	1,028
CD-12	3.71	658	0.19	8,707
CD-13	14.55	39	2.2	1,012
CD-14	6.29	242	4.87	368
CD-15	1.12	7,441	5.42	293
CD-16	6.22	224	1.34	4,455
CD-17	0.88	9,415	1	9,300
CD-18	8.86	104	6.24	129
CD-19	11.32	69	3.3	594
CD-20				
CD-21	6.75	205	1.8	2,945
CD-22	3.5	773	8.97	113
CD-23	3.93	618	7.9	150
CD-24	6.94	194	8.17	139
CD-25	3.55	744	7.53	160
CD-26	2.82	1,126	8.93	103
CD-27	5.61	300	4.93	349
CD-28				
CD-29	10.32	81	3.32	696
CD-30				
CD-31	7.45	168	5.79	280
CD-32	7.65	163	2.83	1,198
CD-33	12.57	55	3.61	450
CD-34	2.87	1,088	8.35	119
CD-35	13.07	54	0.76	7,393
CD-36	11.14	73	4.63	316
Total		30,289		70,627

Table Appendix C.1: sample size needed in independent probability surveys to obtain the levels of accuracy observed. Errors are reported in percentage points. The table is empty where the election was not contested. A maximum effective sample-size bound is set to  $n = 9600$ . Note: decimal changes in area estimates due to simulation uncertainty can lead to increases or decreases of thousands of effective units, hence caution in interpreting these numbers is needed. Some degree of simulation uncertainty persists, as the number of simulations that we used to characterize uncertainty is relatively small - only 500; this was capped because we calculated turnout at the individual level, and it was exceedingly computationally expensive to produce simulations for each of the 13 million Registered voters. Note that inference over average error, average bias at the category level, and success in calling area-level results is robust to this uncertainty.

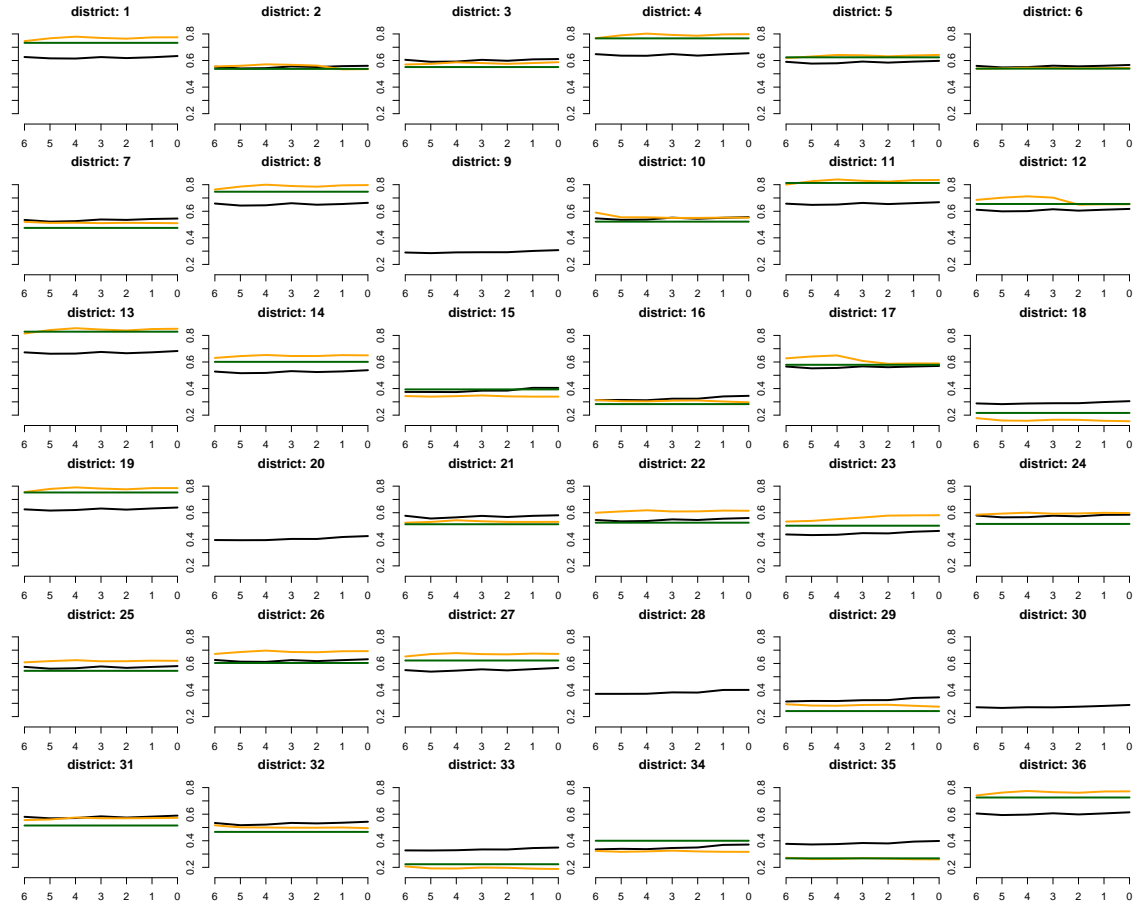


Figure Appendix C.1: Congressional District Forecasts: FiveThirtyEight (orange lines) and Digital Vote (black lines) against election results (in green) over the campaign.

## Appendix D. Information Test

	<i>Dependent variable:</i>		
	2018 Election Result (1)	Change from Trump 2016 (2)	Change from Trump 2016 (3)
538 Predictions	0.868*** (0.097)		
Digital Vote	0.002 (0.172)	−0.333*** (0.075)	
2016 Trump Vote		1.223*** (0.049)	
538 Pred. Change from Trump 2016			0.045 (0.074)
Digital Vote Pred. Change from Trump 2016			−0.095** (0.044)
Constant	0.049 (0.046)	0.042** (0.017)	−0.017*** (0.003)
Observations	33	33	33
R <sup>2</sup>	0.962	0.994	0.188
Adjusted R <sup>2</sup>	0.959	0.993	0.134
Residual Std. Error (df = 30)	0.034	0.014	0.016
F Statistic (df = 2; 30)	376.969***	2,317.410***	3.480**
<i>Note:</i>		*p<0.1; **p<0.05; ***p<0.01	

Table Appendix D.1: Regression table confronting the question: does the Digital Vote add any novel insights to polls based forecasts? The insignificant coefficient in model (1) suggests no new information is present. No new information is added relative to the 2016 Trump Vote (2). No new information on the change since 2016 is present at the district level (3).



## References

- [1] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019.
- [2] T. Enamorado, B. Fifield, K. Imai, Using a probabilistic model to assist merging of large-scale administrative records, 2018.
- [3] W. Wang, D. Rothschild, S. Goel, A. Gelman, Forecasting elections with non-representative polls, *International Journal of Forecasting* 31 (2015) 980–991.
- [4] M. K. Buttice, B. Highton, How does multilevel regression and poststratification perform with conventional national surveys?, *Political Analysis* 21 (2013) 449–467.
- [5] Y. Ghitza, A. Gelman, Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups, *American Journal of Political Science* 57 (2013) 762–776.
- [6] B. E. Lauderdale, D. Bailey, Y. J. Blumenau, D. Rivers, Model-Based Pre-Election Polling for National and Sub-National Outcomes in the US and UK, Technical Report, Working paper, 2017.
- [7] J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, A. Ziegler, Probability machines, *Methods of Information in Medicine* 51 (2012) 74–81.
- [8] M. N. Wright, A. Ziegler, Ranger: a fast implementation of random forests for high dimensional data in c++ and r, *arXiv preprint arXiv:1508.04409* (2015).
- [9] N. Silver, Forecasting the race for the house, 2018. [Online; posted 06-November-2018].
- [10] N. Silver, Forecasting the race for the senate, 2018. [Online; posted 06-November-2018].

- [11] D. K. Park, A. Gelman, J. Bafumi, Bayesian multilevel estimation with poststratification: State-level estimates from national polls, *Political Analysis* 12 (2004) 375–385.
- [12] M. Downes, L. C. Gurrin, D. R. English, J. Pirkis, D. Currier, M. J. Spittal, J. B. Carlin, Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples, *American Journal of Epidemiology* 187 (2018) 1780–1790.
- [13] C. P. Kiewiet de Jonge, G. Langer, S. Sinozich, Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression with Poststratification (MRP), *Public Opinion Quarterly* 82 (2018) 419–446.
- [14] X. Zhang, J. B. Holt, S. Yun, H. Lu, K. J. Greenlund, J. B. Croft, Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system, *American Journal of Epidemiology* 182 (2015) 127–137.
- [15] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, *Science* 348 (2015) 1130–1132.
- [16] X. Zhang, Social media popularity and election results: A study of the 2016 taiwanese general election, *PLOS ONE* 13 (2018) 1–17.
- [17] P. Barbera, Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data, *Political Analysis* 23 (2015) 7691.
- [18] J. B. Kristensen, T. Albrechtsen, E. Dahl-Nielsen, M. Jensen, M. Skovrind, T. Bornakke, Parsimonious data: How a single facebook like predicts voting behavior in multiparty systems, *PLOS ONE* 12 (2017) 1–12.
- [19] S. K. Thompson, Sample size for estimating multinomial proportions, *The American Statistician* 41 (1987) 42–46.

- [20] Igielnik, Ruth and Scott Keeter and Rachel Weisel, Commercial Voter Files and the Study of U.S. Politics, Technical Report, Pew Research Center, 2018.
- [21] O. Keyes, A human name parser for r, <https://github.com/Ironholds/humaniformat>, 2016.
- [22] R. Bond, S. Messing, Quantifying social medias political space: Estimating ideology from publicly revealed preferences on facebook, *American Political Science Review* 109 (2015) 6278.
- [23] R. M. Duch, R. Stevenson, *The Economic Vote: How Political and Economic Institutions Condition Election Results*, Cambridge University Press, Cambridge, 2008.
- [24] M. S. Lewis-Beck, W. G. Jacoby, H. Norpoth, H. F. Weisberg, *The American Voter Revisited*, University of Michigan Press, 2008.
- [25] D. J. Stekhoven, P. Bühlmann, Missforestnon-parametric missing value imputation for mixed-type data, *Bioinformatics* 28 (2011) 112–118.
- [26] S. Janitza, E. Celik, A.-L. Boulesteix, A computationally fast variable importance test for random forests for high-dimensional data, *Advances in Data Analysis and Classification* (2015) 1–31.
- [27] Keeter, Scott and Igielnik, Ruth and Rachel Weisel, Can Likely Voter Modes be Improved? Evidence from the 2014 U.S. House Elections, Technical Report, Pew Research Center, 2016.
- [28] B. C. Burden, Voter turnout and the national election studies, *Political Analysis* 8 (2000) 389–398.
- [29] M. P. McDonald, The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics, *Public Opinion Quarterly* 71 (2007) 588–602.

- [30] Presser, Stanley and Traugott, Michael W. and Santa Traugott, Vote ‘Over’ Reporting in Surveys: The Records or the Respondents?, Technical Report, ANESTechnical Report Series: 010157, 1992.
- [31] B. Lu, Constructing Prediction Intervals for Random Forests, Ph.D. thesis, Pomona College, 2017.
- [32] FiveThirtyEight latest polls, <https://projects.fivethirtyeight.com/polls/senate/texas/>, 2018. Accessed: 2019-01-19.
- [33] A. Graefe, J. S. Armstrong, R. J. Jones Jr, A. G. Cuzán, Combining forecasts: An application to elections, *International Journal of Forecasting* 30 (2014) 43–54.
- [34] D. Denisko, M. M. Hoffman, Classification and interaction in random forests, *Proceedings of the National Academy of Sciences* 115 (2018) 1690–1692.
- [35] C. Hanretty, B. E. Lauderdale, N. Vivyan, Comparing strategies for estimating constituency opinion from national survey samples, *Political Science Research and Methods* 6 (2018) 571591.
- [36] M. K. Buttice, B. Highton, How does multilevel regression and poststratification perform with conventional national surveys?, *Political Analysis* 21 (2013) 449–467.
- [37] J. R. Lax, J. H. Phillips, How should we estimate sub-national opinion using mrp? preliminary findings and recommendations, in: annual meeting of the Midwest Political Science Association, Chicago.
- [38] Y. Ghitza, A. Gelman, Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups, *American Journal of Political Science* 57 (2013) 762–776.