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
[*]

Abstract

Individuals make an increasingly voluminous number of “digital” choices or decisions on a daily basis. We propose a novel MRP estimation strategy that combines samples of these digital traces with a population frame with extensive individual-level socio-economic data in order to generate area forecasts of the outcome variable of interest. In our example, we forecast the two-party vote share for Democrats and Republicans in the 2018 Texas congressional district elections (all 36 districts) and the senate seat election. Our implementation assumes we can observe, and sample, individuals signaling their preference by favoring one virtual location over another. In our case, visiting a Democrat versus Republican Facebook page during the election campaign. We demonstrate that a relatively large 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 eight weeks preceding the mid-term elections we generate vote share forecasts for all 36 congressional seat contests and for the senate race. The forecasts do not use any survey results as input. Nevertheless, they generate vote share forecasts that are accurate when compared to the actual outcomes.

[*] Paper prepared for presentation at the 2019 Annual Asian Political Methodology Meeting, Doshisha University, Kyoto, Japan January 5-6, 2019

Introduction

Observing and recording a sample of social media activity can be the basis for precise predictions about behavior in the population. We demonstrate this is the case with forecasts for the 2018 mid-term elections in the State of Texas. The population frame 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 population frame. We then implement a novel random forest estimation model to aggregate individual preferences in each cell and generate predicted vote share for the parties in each congressional district. The essay provides a number of important insights into how digital trace can inform models predicting human behavior.

There are two important features of our estimation strategy, based on prediction and post-stratification approach. First, the point of departure is a population frame that has individual records for a significant proportion of the population. In our election prediction example, 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 population frame reflect the individual characteristics that we believe determine the behavior of interest – in our example, vote choice. By generating a sufficiently large sample of virtual partisan decisions we can 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, a young black Republican auto-worker could be matched to an appropriate cell in our State of Texas population frame. We use the `R`[12] package `fastLink`[4] to match our virtual sample 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. For many of the individuals observed visiting either a Republican or Democrat candidate's Facebook page, we can determine whether they voted in either a Republican or Democratic primary. We find a very strong correlation between the individuals' digital partisanship and the partisan primary in which they voted, as well as their partisanship status on their voter registration record.

The partisan composition of individual cells in our sampling frame (young, black, college graduates, for example) is informative to the extent that the likelihood of interacting with political content on social media is independent of voting preferences for major voting groups. We confirm this independence with a comparison of the social media partisanship of our virtual sample cells to the recorded partisanship of our population cells. This independence allows us to generate an informative distribution of Republican and Democratic partisans within each cell of the state population frame. More generally, it suggests that samples of these digital traces provide a robust and representative measure of revealed preference.

For each cell in the population frame 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)[17, 2, 5, 8]. We employ a Probability Machine[10] which is a random-forest based algorithm implemented via the `ranger`[18] 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.

Finally, and most importantly, the pre-election vote share predictions for both senate and congressional districts performed very well when benchmarked against actual election results. Our predictions, again based entirely on digital trace data along with historical statistics, also benched very well against the pre-election forecasts of five-thirty-eight[13, 14] that are based on aggregated public opinion surveys.

The essay proceeds as follows: We first describe the virtual sampling strategy – how we generated a large sample of partisan digital traces of eligible Texas voters. This is followed by a discussion of 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, finally we end with a presentation of the eight week-by-week election forecasts that we generated over the course of the mid-term election period.

The Virtual Sample and Population Frames

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[3].

The method has four fundamental components: 1) a population frame 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 predicts 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).

These predicted outcomes for each cell are accurate to the extent that we get right the estimated coefficients from the individual-level model (assuming we have the correct population frame). Current practice is to estimate these individual-level coefficients with relatively large samples of the population of interest[11]. Some evidence suggests that the performance of post-stratification estimation varies significantly depending on the features of these survey samples [2].

On the other hand, there is evidence that post-stratification performs relatively well with unconventional sampling frames. [17] used MRP to forecast the 2012 election outcomes based on an unconventional convenience sample. They administered a brief 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-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. In fact, individual-level estimates from highly skewed sampling frames are the basis for many MRP studies [3, 19]. We note that a large part of their success lies in the size of the sample: under MRP estimation we still need to be powered to accurately estimate vote likelihoods of sub-categories under the ignorability assumption; a smaller sample captured in the same way may not have succeeded in correctly calibrating rare sub-categories of voters.

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. These Facebook users are categorized as partisans based on the pages they visit and then are added to our virtual convenience sample. Post-stratification estimation is then based on this partisan virtual sample.

The 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 a competing Instagram campaign. 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 50 days 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.

Our Texas population frame has $25,920$ cells. For each of the 36 Texas congressional election districts we disaggregate the population by registered partisanship (two 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 720 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 partisans, 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. An important assumption we make is that the conditional propensity to interact with social media is independent of partisanship.[1]. Hence, were we to observe twice as many younger white women liking a Democrat versus a Republican Facebook page is not a function of younger white female's having a higher propensity to interact on social media than is the case for younger white female Republicans.

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 most relevant categories in the population – in our case, voters. To aid us in tuning sample size, we assume a sampling distribution for cell counts of partisan voters of the following form:

$$n_{gr} \sim \text{Multinomial}(p_{1,r=1}, ..., p_{G,r=1}, p_{1,r=0}, ..., p_{G,r=0}, N); \tag{1}$$

where n_{gr} represents the sampled counts of voters belonging to cell g, where $g = 1, ..., G$, 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 Thomposn[16]. We need to specify the nature of our population frame; in our case, indicating the relevant socio-economic categories for predicting vote choice.

We rely on a rich literature on voting behavior in Texas, historical exit-polls and practical resource considerations to define the cells in the sampling frame. We set the sample size to ensure that we can have sample frame cells that represent at least 2.5 percent of the population. 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 sampling theory to inform a-priori expectations regarding sample-size for desired power, this is still a non-probability convenience sample due to the practical sampling strategy and nature of the digital locations.

Our sampling strategy is explicitly dynamic – we aimed 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 exactly 7 weeks before election-day, hence monitoring the period starting on the 18^{th} of September and ending on the 6^{th} of November. Assuming again 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. In practice, the effective weekly sample size must be 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. Our estimates of stable partisans will be precise.

[1] Benchmarking the smoothed cell preferences with results from a probability sample will reveal if this assumption is too strong, in which case discrepancies in the distribution of partisanship over cells will be non-trivial.

We adopted explicit steps to avoid introducing bias as a result of systematic day-of-the-week effects on clustering of preferences[2]. We spread the collection of partisan voters evenly over the days of the week, and enforced the quota at the end of the week via re-sampling at random from the pool of partisan voters captured, until the partisan proportions are as designed.[3]

To account a-priori for potential baseline differences in partisanship on social media (i.e. social media activity is systematically conditional on partisanship) we impose a 50 percent weekly party-quota. We try to ensure exactly half of our weekly sample is Republican, and the other half Democrat. This quota reflects the a-priori belief that the state is a "toss-up". We expect this assumption to be roughly accurate for the Texas experiment but recognize it may need adjustments when more complex dynamics are in play.

The number of users captured by the original procedure per week is 5 per 68 pages per day, amounting to a maximum number of sampled users per week of $2,380$. 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 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 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.[4]

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 consistently and severely out of the norm for the given cell.

In total, we sampled $15,683$ digital traces. Prior to re-sampling, 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 step, and one this in most cases the most challenging, is to match the virtual sample to the population. The potential applications of this MRP strategy are constrained by this requirement that the digital trace be matched with, ideally, records of individuals in the population.

[2] 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.

[3] The Appendix has Routine 1 that describes the data collection steps in detail.

[4] Most pages we sample from correspond to a congressional districts. 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 considered using congressional district as a variable in our likelihood estimation; but if any district-candidate page is systematically less active than that of their district rival, district co-variates would capture artificial variation due to heterogeneous page-activity rates, and activity could be confounded with voting preferences.

Matching Sample to the Population Frame

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 sample 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. In order to match these digital preferences to the Texas population frame we need richer socio-demographic information on the individuals. There are various 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 could imagine, for example, conducting conventional, obtrusive, surveys to obtain this information.

It will occasionally be the case though that there exists reasonably comprehensive, and accessible, lists of individuals in the population that contain this augmented socio-demographic information. In our case, we get these richer profiles by matching our social media sample to the state-wide Texas L2 voter registration file.[5] 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 filter non-registered voters and generate turnout weights based on historical voting patterns.

We obtain three variables from the Facebook individual profiles that can match to the L2 voter history file. Name is one of them. We use the `R`[12] package `humaniformat`[7] 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 our voter list counterparts. Secondly, we want to match the 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 these inconsistencies and mistakes in data entry 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`[4] 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 deals with missing data by imputing them from the posterior distribution, a characteristic 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.*

[5] L2 is a nonpartisan supplier of voter data; see `https://www.l2political.com` for details.

Threshold	0.75	0.85	0.9	0.95	0.99
Match Count	8739	8394	8384	8283	7322
Match Rate (%)	54.59	52.89	52.84	52.24	46.31
FDR (%)	2.03	1.18	1.16	1.08	0.81
FNR (%)	99.64	99.65	99.65	99.66	99.7

Table 1: `fastLink` output summary table.

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 whilst 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.

Table 1 displays the match rate under three different thresholds; it further reports the estimated False Discovery Rate (FDR) and False Negative Rate (FNR). The FDR is the proportion of false matches having overall posterior matching probability higher than the given threshold; the FNR is the proportion of true matches whose match probability falls below the threshold. The model is optimized to minimize FDR. FNR is exceedingly high but constant across thresholds, reflecting the sheer number of potential matches. Holding the FDR close to zero is of greater importance to us than minimizing the FNR – given our social media sample is quite small, we cannot afford to have the matching introduce further noise.

	Dem. Digital Trace	Rep. Digital Trace
Registered Democrat	3861	868
Registered Independent	217	182
Registered Republican	575	2575
Dem. Primary Voter	2606	253
Rep. Primary Voter	278	1777

Table 2: Relationship between digital trace and objective measures of partisanship; these can be interpreted as two stacked contingency tables. Discrepancies between the sum of these numbers and the total matched sample size are due to missing data in the measures of partisanship. It is relevant to count all digital traces, as opposed to individuals, as each individual can leave multiple, disagreeing traces.

From our social media sample of $15,683$ digital traces, we successfully match $8,278$ traces to an entry in the voter-registration – this is a success rate of 52.8%. To assess our assumption that digital traces are an expression of political preferences, we check whether the digital traces match a) the stated partisanship of the individuals matches to the voter file; and b) the partisanship of the primary these individuals voted in, if they cast a primary ballot. As Table 2 indicates the probability that a digital trace is Republican, conditional on its user being a registered Republican: P(Dig. Rep.|Reg. Rep.) = 0.82. The probability that a digital trace is Democrat, given the user is a registered Democrat: P(Dig. Dem.|Reg. Dem.) = 0.82. Similar quantities using primary turnout give comparable results: P(Dig. Rep.|Rep. Primary) = 0.86; P(Dem. Rep.|Dem. Primary) = 0.91. There is a high correlation between digital traces and political preferences. And the constant registration-conditional probabilities across parties suggest a stable relationship between digital preferences and stated preferences. We are very comfortable concluding that digital traces represent revealed preferences.

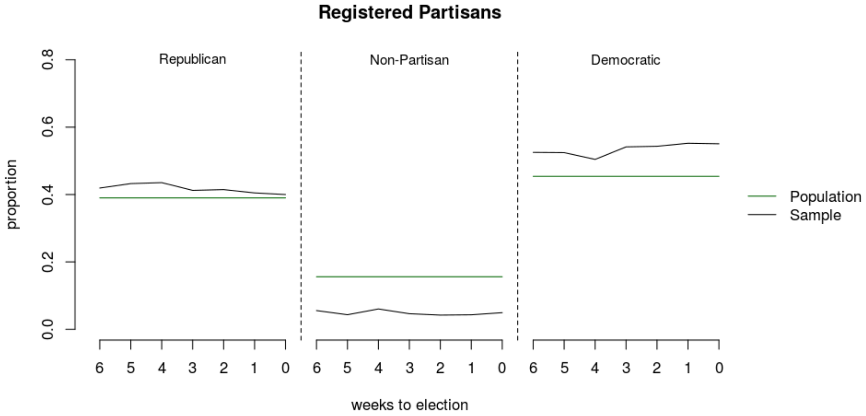

Figure 1: Population v. Sample comparison: partisanship.

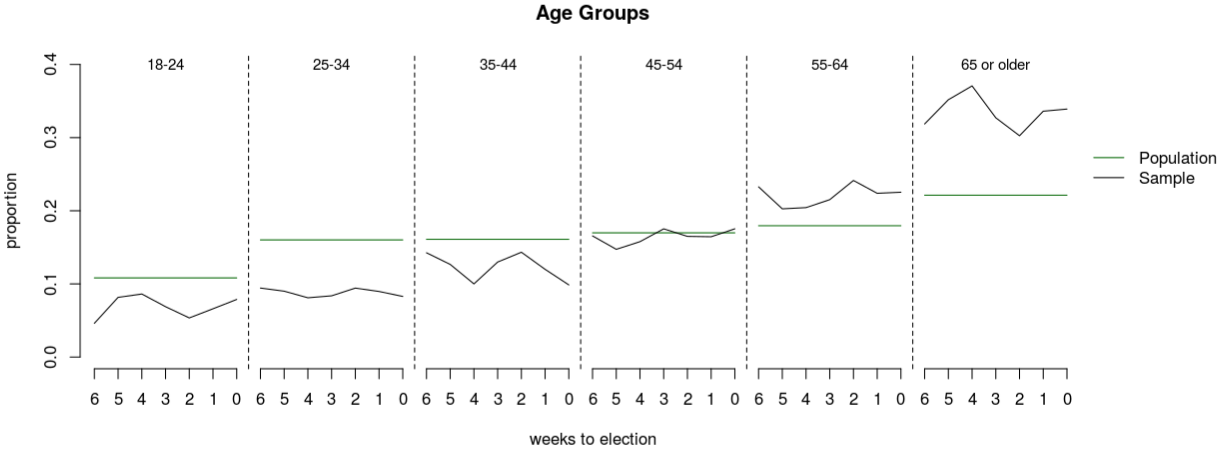

Figure 2: Population v. Sample comparison: age.

At the end of the matching process we re-sample digital traces from each week to meet the quota defined in our sampling strategy. Our final sample size is 9,306 digital traces. This is slightly below the Thompson number. Holding probability constant at 0.9, we are powered (assuming equal sample size across weeks), at the weekly level, to estimate voter categories that are larger than 2.75 percent of the total population.

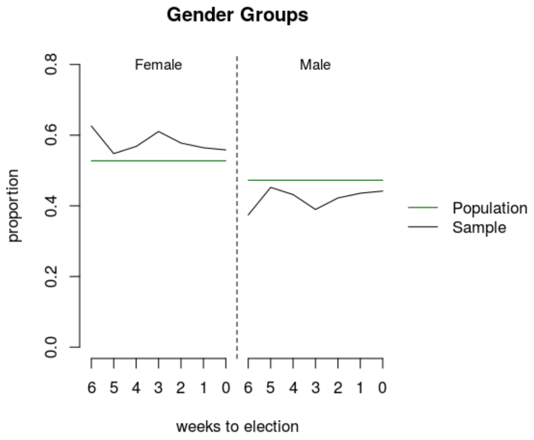

Figure 3: Population v. Sample comparison: gender.

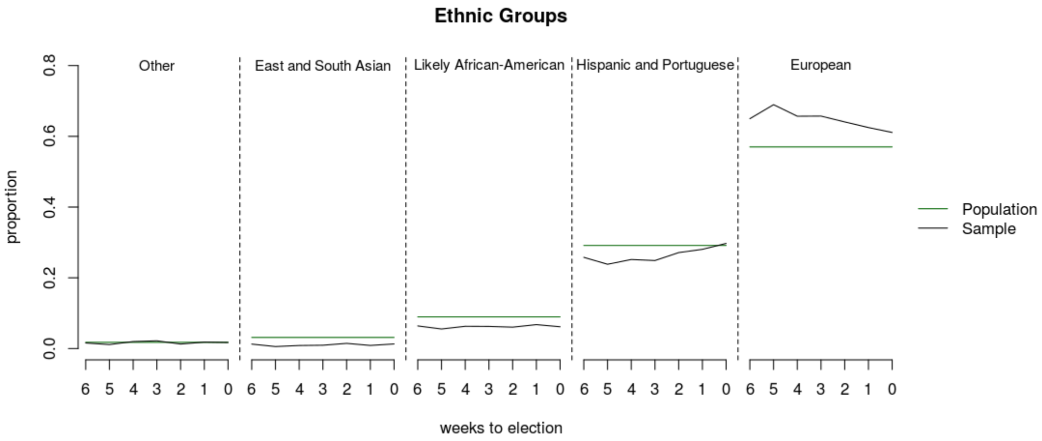

Figure 4: Population v. Sample comparison: ethnicity.

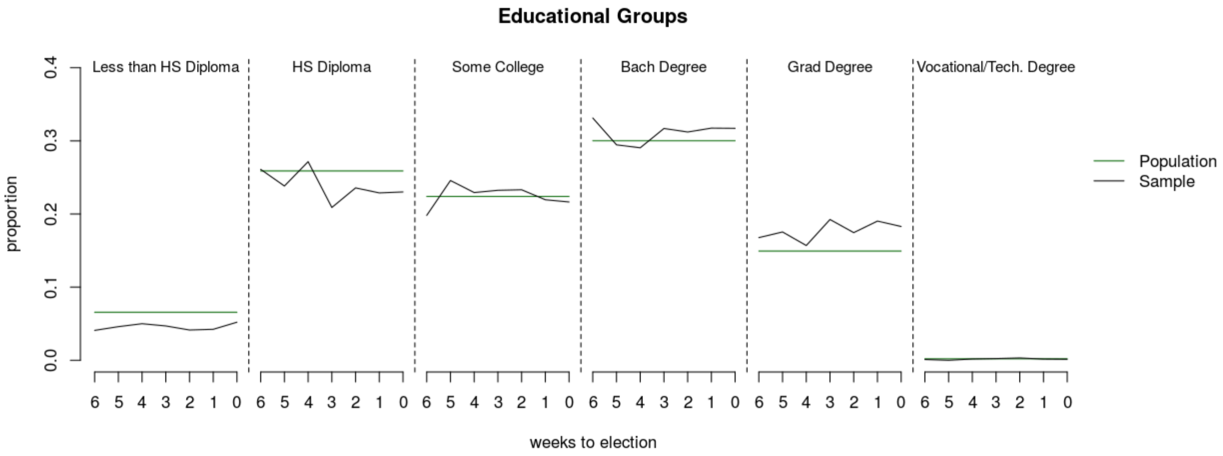

Figure 5: Population v. Sample comparison: highest level of education.

Figures 1 to 5 compare the socio-demographic profile of our matched sample to that of the population at large. With respect to *Registered Partisanship*, we essentially overestimate the percentage of Democratic and Republicans in the population and we understate the percentage of Independents. This is an understandable result of sampling from partisan Facebook pages. With respect to *Age*, we underestimate the young and overestimate the old. On *Gender* we slightly under-represent men and over-represent women. The *Ethnic* composition of 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.

This is a large convenience sample of individuals visiting partisan Facebook pages. A critical element of our design is matching these individuals to a population frame – in our case all Texas registered voters. The matching allows us to characterize 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 large virtual samples can be quite representative of the overall population.

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[17, 2, 5, 8]. 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. The novelty of our estimation strategy is that we do not rely on survey responses to estimate our model of vote preference. The outcome of interest, 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 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.

Variable Selection and Imputations

We select just under 50 variables (see Table 2 in the Appendix), based on two primary considerations: a) they are expected to be good predictors of turnout or vote choice; b) they provide useful information for imputing missing variables. For each h registered voter we have $\boldsymbol{x}_h$. Missing values in $\boldsymbol{X}$ are imputed with a random-forest multiple-imputation strategy implemented via the packages `ranger`[18] and `missForest`[15]. 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 $\boldsymbol{X}^{\mathrm{I}}$ with voter-specific characteristics as $\boldsymbol{x}_h^{\mathrm{I}}$.

From $\boldsymbol{X}^{\mathrm{I}}$ we derive two datasets: $\boldsymbol{X}^{\mathrm{I}^+}$, indexed by expanded voter identifier h^+ and $\boldsymbol{Z}^{\mathrm{I}}$, indexed by h as above. The former is an expanded version of the completed voter sub-space, deployed for estimating turnout probabilities; it is constructed by using the covariate "voted in year ____" as outcome variable, warping $\boldsymbol{X}^{\mathrm{I}}$ into *long* format accordingly.

Unlike vote preference, we do not have a digital trace measure of turnout likelihood. To estimate voter turnout probabilities we therefore rely on individual level historical voting behaviour from the voter-registration file. 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.

$\boldsymbol{Z}^{\mathrm{I}}$ is of the same length as its mother, but the covariates which compose it are reduced to the subset of size m which uniquely identifies voter-categories of interest. We define the voter category C_g, for categories $g = 1, ..., G$ as a unique realization of the set of variables which compose $\boldsymbol{Z}^{\mathrm{I}}$, i.e. $C_g = \{Z_1 = z_1, ..., Z_m = z_m\}$; then for a voter h in the registry, $h \in g$ if $\boldsymbol{z}_h^{\mathrm{I}} = C_g$. Table 2 in the technical 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 (remember from Thomposn[16] we are powered to capture groups of voters which make up roughly 2.5% of the registered electorate). The resulting voting categories are then used to characterize the cells of our population frame.

As was mentioned in the sampling strategy, the voter categories $\boldsymbol{C}$ 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 $\boldsymbol{C}$ 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 $d = 1, ..., 36$) 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.

Likelihood Prediction Models

Throughout this section we omit the imputation overscript "I" and simply refer to relevant completed datasets as $\boldsymbol{X}$ and $\boldsymbol{Z}$. We want to estimate the joint probability of an individual in category g voting for the Republican candidate and turning out on election-day; we decompose the problem similarly to Lauderdale et al.[8]:

$$\mathrm{P}_g\left(R = 1, T = 1|C\right) = \mathrm{P}_g\left(R = 1|T = 1, C\right) \times \mathrm{P}_g\left(T = 1|C\right) \tag{2}$$

. We again rely on the `ranger` package to implement a probability machine[10] composed of B^T trees. 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. Using random forests we improve the previous MRP methodology on two fronts. First, the forest estimation outputs the best[6] non-linear function of its inputs without us imposing a-priori functional specifications. If the input is not *important*[6] 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.

Voter Turnout Estimation

Turnout behaviour is estimated with two probability distributions: i) $\mathrm{P}_h\left(T = 1|\boldsymbol{x}\right)$ for individuals in the completed voter file $\boldsymbol{X}$, where T is a random variable taking value 1 if the individual casts a ballot on election-day and 0 otherwise; ii) its average over the relevant voter-categories used to predict vote-choice likelihood, $\mathrm{P}_g\left(T = 1|C\right)$. The former point estimate, P_h, is used as observation-weight in the estimation of the vote-choice likelihood, for each of our matched social media voters. The latter, P_g, is used to calculate the category-level probability of turning out on election day, which will be needed to calculate Equation 2. We train a random forest on the expanded turnout dataset $\boldsymbol{X}^+$; the forest's estimate of the probability of turning out on election day conditional on voter characteristics and being registered is defined as follows:

$$\hat{\mathrm{P}}_{h^+}\left(T = 1|\boldsymbol{x}^+\right) = \varphi^T\left(\boldsymbol{x}^+_{h^+}\right) = \frac{1}{B^T}\sum_b^{B^T}\tau^T_b\left(\boldsymbol{x}^+_{h^+}\right); \tag{3}$$

where φ^T represents the point estimate of a probability machine trained for prediction of turnout probabilities, and whose value is the average of B^T probability trees τ^T_b. 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:

$$\hat{\mathrm{P}}_h\left(T = 1|\boldsymbol{x}\right) = \varphi^T\left(\boldsymbol{x}_h\right); \tag{4}$$

We first extract the turnout probabilities of voters in our matched social media sample $\hat{\mathrm{P}}_s\left(T = 1|\boldsymbol{x}\right)$ indexed by $s = 1, ..., S$ where $s \subset h$; this quantity will be used as the observation-weights in our vote-choice model, effectively conditioning that distribution on turnout. Second, we average average across the voter categories identified by $\boldsymbol{C}$ to obtain category-level estimates of turnout probabilities, as follows:

$$\mathrm{P}_g\left(T = 1|C\right) = \frac{1}{\sum_{h\in g}\mathbb{1}\left(\boldsymbol{z}_h = C_g\right)}\sum_{h\in g\forall\boldsymbol{z}_h = C_g;\boldsymbol{z}_h\in\boldsymbol{x}_h}\mathrm{P}_h\left(T = 1|\boldsymbol{x}\right); \tag{5}$$

[6] According to minimization procedure of an out-of-bag prediction error

where the outcome quantity is the average across a number of simulations from the predictive distribution of $\mathrm{P}_h\left(T=1|\boldsymbol{x}\right)$.

We calculate the empirical error distribution using the MSPE1 procedure from Lu[9]; 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:

$$\mathrm{P}_h\left(T=1|\boldsymbol{x}\right) \sim \mathrm{N}\left(\varphi^T\left(\boldsymbol{x}_h^{\mathrm{I}}\right), (\hat{\sigma}^T_{\mathrm{RMSE1}})^2\right). \tag{6}$$

The occurrence of negative probabilities (or probabilities above 1) as we simulate from this predictive distribution to characterize the variance is not a huge issue because aggregating to obtain area estimates tends to shrink the distribution of the area estimates to fit within the probability space.

Estimating Vote Probabilities

To estimate vote-choice probabilities we specify the outcome variable as the partisan digital trace R_s of voters $s = 1, ..., S$ where $R_s = 1$ if the trace is Republican and 0 if Democrat. 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:

$$\hat{\mathrm{P}}_s\left(R=1|T=1,\boldsymbol{z}\right) = \varphi^R\left(\boldsymbol{z}_s|\hat{\mathrm{P}}_s\left(T=1|\boldsymbol{x}\right)\right) = \frac{1}{B^R}\sum_b^{B^R}\tau_b^R\left(\boldsymbol{z}_s|\hat{\mathrm{P}}_s\left(T=1|\boldsymbol{x}\right)\right); \tag{7}$$

Here, $\boldsymbol{Z}$ includes an identifier for each week to election $W = w$ for $w = 7, ..., 0$; 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. Accordingly, we expand our definition of categories $\boldsymbol{C}$ to account these new identifiers. The category-conditional turnout estimated previously is assumed to be constant across W and and L. The trained forest is then used to provide category-level predictions; for category g such that $s \in g$ if $\boldsymbol{z}_s = C_g$:

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

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:

$$\mathrm{P}_g\left(R=1|T=1,C\right) \sim \mathrm{N}\left(\varphi^R\left(C_g\right), (\hat{\sigma}^R_{\mathrm{RMSE1}})^2\right); \tag{9}$$

In this section we propose a novel approach to the model estimation stage of classic MRP area estimation.

First, the outcome variable in our model, vote preference, is observed, unobtrusively – in our illustration, we do not rely on self-reported vote preferences. 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 estimated outcome probabilities 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 provides, in our example, the two quantities needed to estimate the cell-level joint probability of voting Republican and turning out, as specified in Equation 2.

Aggregation: Vote and Seat Predictions

Ultimately, of course, we are interested in estimating values for out outcome variable for the areas of interest. In our example, this would be the congressional vote predictions for each of the 36 congressional districts and the senate vote predictions for the overall 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: $Q_g = \sum_h \mathbb{1}(\boldsymbol{z}_h = C_g)$. 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:

$$V^R_{gw} = \frac{\sum_g \mathrm{P}_g\left(R=1, T=1 | C, L=1, W=w\right) \times Q_g}{\sum_g \mathrm{P}_g\left(T=1 | C, L=1, W=w\right) \times Q_g}. \tag{10}$$

At the district level we count the number of voters in the intersection between a given category and a given district: $q_{gd} = \sum_h \mathbb{1}(\boldsymbol{z}_h = C_g \cap D_h = d)$. The district level estimates are the product of the following calculation:

$$v^R_{gdw} = \frac{\sum_g \mathrm{P}_g\left(R=1, T=1 | C, L=0, W=w\right) \times q_{gd}}{\sum_g \mathrm{P}_g\left(T=1 | C, L=0, W=w\right) \times q_{gd}}. \tag{11}$$

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

Results: Forecast Election Results

Area forecasts that use prediction and post-stratification methods rely on a sample of the population frame on which to base the category predictions. Rather than a conventional survey of the population, we propose a virtual sample that measures, unobtrusively, the outcome variable matched with a population frame that provides individual-level measures that, again, are not based in self-reporting. In our case we employ an unobtrusive sample of Facebook users who visit partisan Facebook pages. Our evaluation of the estimation strategy is based on the 2018 Texas congressional elections. Using this virtual sampling method we generated weekly-updated forecasts of the vote shares for Republican and Democrat candidates in each of the state's thirty-six congressional districts. Similar predictions were generated for the two leading Senatorial

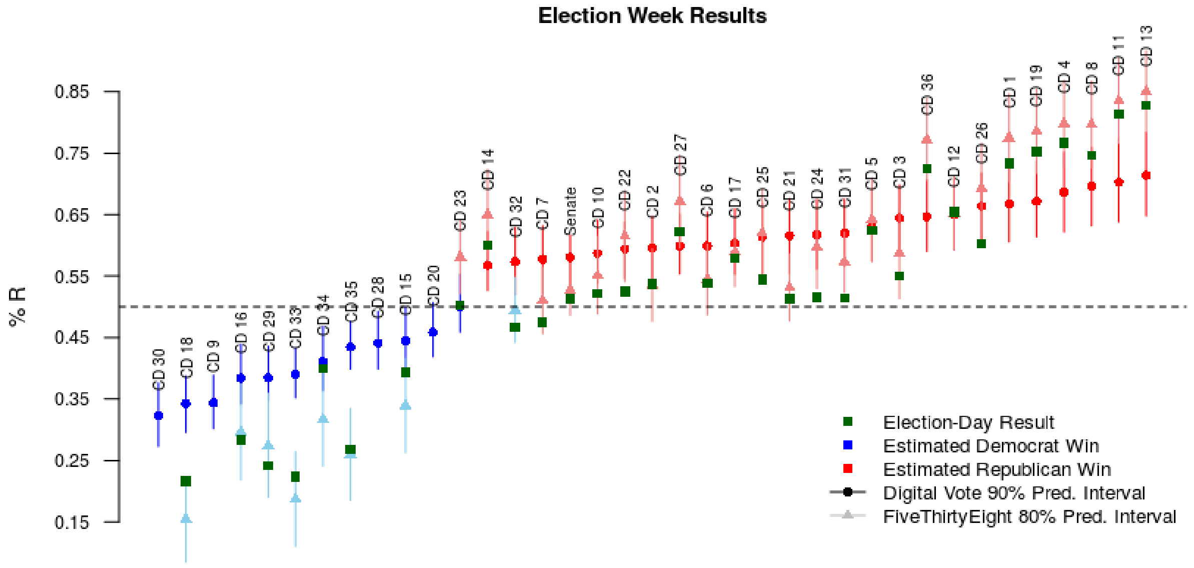

Figure 6: Estimates of support for the Republicans, by area of interest; a comparison with `FiveThirtyEight.com` and actual election results is provided.

candidates. These are what we have labeled “Digital Vote” predictions.

Figure 6 presents the Digital Vote Senate and Congressional predictions for the election week period. Thirty-three of the 36 congressional districts were not particularly competitive and in these easy cases our predictions match the actual election outcome. There were three particularly competitive Districts: District 7 and District 32 went for the Democrats and we predicted Republican wins. And in District 23, we correctly characterized it as a toss-up and correctly predicted on election week a Republican win.

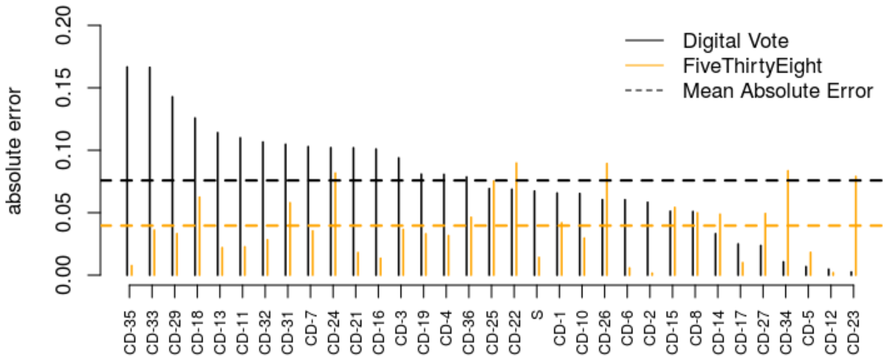

Figure 7: Mean Absolute Error: FiveThirtyEight and Digital Vote

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 proprietary aggregation of the latest public opinion polls. The site puts out several different forecasts, which aggregate a variety of information. We choose the "light" version as it has no other information beyond polls, and hence provides a more direct comparison to our method, which seeks to emulate an opinion poll using a digital sample.

The election week comparisons in Figure 6 suggest that the FiveThirtyEight aggregated results perform somewhat better than the Digital Vote. Note that FiveThirtyEight correctly forecast District 7 which was not the case for Digital Vote. Also, FiveThirtyEight tends to have estimated vote shares that are closer to actual outcomes.

Figures 9 in the Appendix plots the forecasts of the Digital Vote forecasts and those of FiveThirtyEight for each of the Texas congressional districts and for the senate race. Conventional survey methods and our "virtually" informed MRP generate remarkably similar results. Many of these districts have large vote margins in favor of one of the parties – and the two approaches always agree in these cases. Figure 7 summarizes the Mean Absolute Error for the two methods over the 36 congressional districts and the senate seat. These are based on the differences between the weekly estimated vote share and actual vote share observed in the election.

There are two important messages in Figure 7. First, overall, MRP forecasts based on our novel virtual sampling strategy have relatively small Mean Absolute Error. Over all 36 districts the mean absolute error is about 7.5 percentage points. Second, 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. Our Digital Vote forecasts, particularly for the least competitive races, tend to gravitate closer to 0.5 than is the case for FiveThirtyEight, symptomatic of attenuation bias. Nevertheless, as we saw above, the two forecasting strategies were in close agreement regarding the ultimate victors in each of the congressional districts.

Of particular interest here is the cost of achieving this level 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 $n = \left(\frac{z_{\alpha/2}(\pi(1-\pi))}{d}\right)^2$ with π set to a point estimate, 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 7, 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. Table 3 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 $730,808$ individuals to achieve FiveThirtyEight's level of accuracy, and $229,646$ individuals to achieve ours. Considering our matched sample was $9,306$ digital traces, via our procedure we have obtained levels of accuracy which are roughly 25 times its weight in individuals sampled.

District	Digital Vote Error	Digital Vote n	FiveThirtyEight Error	FiveThirtyEight n
Senate	6.73	207	1.42	4,782
CD-1	6.57	198	4.19	382
CD-2	5.83	272	0.15	418,486
CD-3	9.38	100	3.65	698
CD-4	8.07	127	3.16	618
CD-5	0.68	19,600	1.82	2,665
CD-6	6.04	253	0.57	29,507
CD-7	10.29	89	3.53	769
CD-8	5.09	314	4.99	250
CD-9				
CD-10	6.53	218	2.96	1,084
CD-11	10.99	66	2.27	1,028
CD-12	0.47	39,658	0.19	234,464
CD-13	11.4	60	2.2	1,012
CD-14	3.31	860	4.87	368
CD-15	5.11	363	5.42	293
CD-16	10.09	89	1.34	4,455
CD-17	2.49	1,478	1	9,389
CD-18	12.58	55	6.24	129
CD-19	8.08	130	3.3	594
CD-20				
CD-21	10.19	88	1.8	2,945
CD-22	6.87	196	8.97	113
CD-23	0.25	154,351	7.9	150
CD-24	10.2	87	8.17	139
CD-25	6.92	190	7.53	160
CD-26	6.04	235	8.93	103
CD-27	2.36	1,651	4.93	349
CD-28				
CD-29	14.28	45	3.32	696
CD-30				
CD-31	10.47	83	5.79	280
CD-32	10.66	83	2.83	1,198
CD-33	16.63	33	3.61	450
CD-34	1.06	8,291	8.35	119
CD-35	16.66	34	0.76	12,817
CD-36	7.86	142	4.63	316
Total		229,646		730,808

Table 3: Hypothetical 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.

We benchmarked the hypothetical sample sizes reflected in the Digital Vote against those of FiveThirtyEight forecasts for the competitive Texas Congressional Districts and Senate election. FiveThirtyEight forecasts in some sense represent a gold standard for benchmarking since they incorporate over 500 public opinion surveys totalling roughly 250,000 respondents for their house forecasts; 450 surveys of over 380,000 respondents for their Senate forecast[1][7]. The polls FiveThirtyEight included in their model which were specific to Texas, conducted during the monitoring period, that in some respect would most strongly inform the FiveThirtyEight Texas congressional and senate seat forecasts, consisted of a total of 22 house surveys of over 11'000 respondents; 26 senate surveys of over 25'000 individuals. Our Texas forecasts are about one-third as precise as the FiveThirtyEight aggregated estimates for the Texas midterm district and senate elections; they are only based on the digital traces of 10,000 Facebook users.

Conclusion

Individuals make an increasingly voluminous number of "digital" choices or decisions on a daily basis. These digital traces can be useful for some forecasting activities. We propose a novel MRP estimation strategy

[7] These figures refer to the number of surveys conducted during our monitoring period.

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. In our example, we forecast vote share for the Democrats and Republicans in the 2018 Texas congressional district elections (all 36 districts) and the senate seat election. 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 demonstrate that a relatively large 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 eight weeks preceding the mid-term elections we generate vote share forecasts for all 36 congressional seat contests and for the senate race. The forecasts do not use any survey results as input. Nevertheless, they generate vote share forecasts that are quite accurate when compared to the actual outcomes.

Our Texas digital trace forecast is an experiment – we are interested in assessing the quality of predictions – in this case voting behavior – that could be generated from revealed digital preferences. As mentioned, we believe the results are promising for election studies but in fact we think there are key features of this exercise that have broader implications. We argue that digital traces are unobtrusive measures of revealed preferences. Hence, for a range of behavioral outcomes of interest, they do not have the widely recognized disadvantages of preferences that are elicited by conventional survey methods. A challenge, though, is 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. And most importantly, there are the predictions. We generate areal estimates of preferences that employ a novel implementation of MRP and machine learning techniques. Here we are modeling unobtrusively-measured digital traces to predict, for the area of interest (in our case U.S. congressional districts), behavior (in this example, vote choice). A sample of digital traces along with this areal estimation method produces predictions of preferences that are highly correlated with actual behavior. We leverage our ability to observe actual vote outcomes in Congressional Districts as a means to validate the vote predictions made for each District.

References

[1] FiveThirtyEight latest polls. https://projects.fivethirtyeight.com/polls/senate/texas/. Accessed: 2019-01-19.

[2] Buttice, M. K., and Highton, B. How does multilevel regression and poststratification perform with conventional national surveys? *Political Analysis 21*, 4 (2013), 449–467.

[3] Downes, M., Gurrin, L. C., English, D. R., Pirkis, J., Currier, D., Spittal, M. J., and Carlin, J. B. Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. *American Journal of Epidemiology 187*, 8 (2018), 1780–1790.

[4] Enamorado, T., Fifield, B., and Imai, K. Using a probabilistic model to assist merging of large-scale administrative records, 2018.

[5] Ghitza, Y., and Gelman, A. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science 57*, 3 (2013), 762–776.

[6] Janitza, S., Celik, E., and Boulesteix, A.-L. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification* (2015), 1–31.

[7] Keyes, O. A human name parser for r. https://github.com/Ironholds/humaniformat, 2016.

[8] Lauderdale, B. E., Bailey, D., Blumenau, Y. J., and Rivers, D. Model-based pre-election polling for national and sub-national outcomes in the us and uk. Tech. rep., Working paper, 2017.

[9] Lu, B. *Constructing Prediction Intervals for Random Forests.* PhD thesis, Pomona College, 2017.

[10] Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., and Ziegler, A. Probability machines. *Methods of Information in Medicine 51*, 01 (2012), 74–81.

[11] Park, D. K., Gelman, A., and Bafumi, J. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis 12*, 4 (2004), 375?385.

[12] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

[13] Silver, N. Forecasting the race for the house, November 2018. [Online; posted 06-November-2018].

[14] Silver, N. Forecasting the race for the senate, November 2018. [Online; posted 06-November-2018].

[15] Stekhoven, D. J., and Bühlmann, P. Missforestnon-parametric missing value imputation for mixed-type data. *Bioinformatics 28*, 1 (2011), 112–118.

[16] Thompson, S. K. Sample size for estimating multinomial proportions. *The American Statistician 41*, 1 (1987), 42–46.

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

[18] Wright, M. N., and Ziegler, A. Ranger: a fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409* (2015).

[19] Zhang, X., Holt, J. B., Yun, S., Lu, H., Greenlund, K. J., and Croft, J. B. 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*, 2 (2015), 127–137.

Appendices

A Sampling

Routine 1 A description of the relevant steps in our sampling routine. The page space ρ 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. We note that in practice we enforce the quota only after we filter the sample via matching with the voter registration file.

```
procedure getUserInfo
    let $\rho$ a vector of relevant social media pages
    let $N$ be a counting operator
    let $M(\rho_{O_d})$ be the subset of users being collected per page, according to daily order $O_d$
    let collect be a function of the input user, with output $z$, the set of relevant characteristics
    let $R(\boldsymbol{z})$ be the subset of Republican partisans, and let $D(\boldsymbol{z})$ be the same for the Democrats

    # FOR EACH WEEK LEFT IN THE CAMPAIGN
    for $w$ in $W, ..., 0$ do

        # FOR EACH DAY WITHIN THE WEEK
        for $d$ in $7, ..., 1$ do

            # SAMPLE AT RANDOM A COLLECTION ORDER
            $O_d$ = sample( from = $1, ..., N(\rho)$ , size = $N(\rho)$ , replacement = FALSE )

            # FOR EACH PAGE IN THE PAGE SPACE
            for $O_d$ in $1, ..., N(\rho)$ do

                # VISIT THE LATEST DAILY POST ON THE PAGE
                goto $\rho_{O_d}$

                # FOR EACH EXPLICITLY PARTISAN USER IN $M(\rho_{O_d})$
                for $i$ in $1, ..., N(M(\rho_{O_d}))$ do

                    # COLLECT INFORMATION FROM THEIR PUBLIC ABOUT PAGE
                    $z_i$ = collect $M_i(\rho_{O_d})$

                end for
            end for
        end for

        # RESAMPLE TO MEET QUOTA
        $d_w = N(R(\boldsymbol{z}_w)) - N(D(\boldsymbol{z}_w))$
        if $d_w > 0$ then
        $z'_w$ = sample ( from = $D(\boldsymbol{z}_w)$, size = $d_w$, replacement = TRUE )
        else
        $z'_w$ = sample ( from = $R(\boldsymbol{z}_w)$, size = $d_w$, replacement = TRUE )
        return $z''_w = (z_w, z'_w)$

    end for

end procedure
```

Page Address	Party	Election Type	District ID	n	% Total
https://www.facebook.com/betoorourke/	Dem	US Senate	0	1521	9.70
https://www.facebook.com/tedcruzpage/	Rep	US Senate	0	1615	10.30
https://www.facebook.com/vote4McKellar	Dem	US House	1	19	0.12
https://www.facebook.com/RepLouieGohmert/	Rep	US House	1	125	0.80
https://www.facebook.com/toddlittonforcongress/	Dem	US House	2	230	1.47
https://www.facebook.com/CrenshawforCongress/	Rep	US House	2	192	1.22
https://www.facebook.com/Lorieburchforcongress/	Dem	US House	3	98	0.62
https://www.facebook.com/VanForTexas/	Rep	US House	3	44	0.28
https://www.facebook.com/KrantzforCongress/	Dem	US House	4	224	1.43
https://www.facebook.com/RepRatcliffe/	Rep	US House	4	168	1.07
https://www.facebook.com/lancegoodenfortexas/	Rep	US House	5	114	0.73
https://www.facebook.com/JanaLynneSanchezforUSCongress/	Dem	US House	6	196	1.25
https://www.facebook.com/wright4congress/	Rep	US House	6	141	0.90
https://www.facebook.com/LizzieForCongress/	Dem	US House	7	786	5.01
https://www.facebook.com/johnculberson/	Rep	US House	7	598	3.81
https://www.facebook.com/stevenforcongress/	Dem	US House	8	26	0.17
https://www.facebook.com/BradyforTexas/	Rep	US House	8	148	0.94
https://www.facebook.com/repalgreen/	Dem	US House	9	105	0.67
https://www.facebook.com/siegelfortexas/	Dem	US House	10	205	1.31
https://www.facebook.com/MichaelMcCaulTX/	Rep	US House	10	106	0.68
https://www.facebook.com/mike.conaway/	Rep	US House	11	136	0.87
https://www.facebook.com/VanessaAdiaTX12/	Dem	US House	12	198	1.26
https://www.facebook.com/RepKayGranger/	Rep	US House	12	111	0.71
https://www.facebook.com/gregsagan2018/	Dem	US House	13	140	0.89
https://www.facebook.com/ThornberryForCongress/	Rep	US House	13	72	0.46
https://www.facebook.com/adrbell/	Dem	US House	14	136	0.87
https://www.facebook.com/WeberForTexas/	Rep	US House	14	41	0.26
https://www.facebook.com/votevicente/	Dem	US House	15	124	0.79
https://www.facebook.com/westley4Congress/	Rep	US House	15	167	1.06
https://www.facebook.com/voteforveronica/	Dem	US House	16	140	0.89
https://www.facebook.com/Seeberger1ForCongress/	Rep	US House	16	204	1.30
https://www.facebook.com/RickKennedyforCongress/	Dem	US House	17	153	0.98
https://www.facebook.com/BillFloresForCongress/	Rep	US House	17	161	1.03
https://www.facebook.com/CongresswomanSheilaJacksonLee/	Dem	US House	18	105	0.67
https://www.facebook.com/Ava-for-Congress-526965147465733/	Rep	US House	18	122	0.78
https://www.facebook.com/miguellevario19/	Dem	US House	19	195	1.24
https://www.facebook.com/JodeyArrington/	Rep	US House	19	142	0.91
https://www.facebook.com/JoaquinCastroTX/	Dem	US House	20	143	0.91
https://www.facebook.com/ChipRoyforCongress/	Dem	US House	21	230	1.47
https://www.facebook.com/KopserforCongress/	Rep	US House	21	235	1.50
https://www.facebook.com/KulkarniforCongress/	Dem	US House	22	225	1.43
https://www.facebook.com/PeteOlsonTX/	Rep	US House	22	126	0.80
https://www.facebook.com/GinaOrtizJones/	Dem	US House	23	755	4.81
https://www.facebook.com/HurdForCongress/	Rep	US House	23	747	4.76
https://www.facebook.com/JanMcDowellDemocrat/	Dem	US House	24	179	1.14
https://www.facebook.com/RepKennyMarchant/	Rep	US House	24	150	0.96
https://www.facebook.com/JulieForTexas/	Dem	US House	25	192	1.22
https://www.facebook.com/RepRogerWilliams/	Rep	US House	25	186	1.19
https://www.facebook.com/LinseyFaganTx/	Dem	US House	26	218	1.39
https://www.facebook.com/michaelcburgess/	Rep	US House	26	175	1.12
https://www.facebook.com/EricHolguin/	Dem	US House	27	222	1.42
https://www.facebook.com/CloudforCongress/	Rep	US House	27	161	1.03
https://www.facebook.com/repcuellar/	Dem	US House	28	53	0.34
https://www.facebook.com/SylviaRGarcia/	Dem	US House	29	147	0.94
https://www.facebook.com/aronoffforcongress/	Rep	US House	29	112	0.71
https://www.facebook.com/CongresswomanEBJtx30/	Dem	US House	30	89	0.57
https://www.facebook.com/MJforTexas/	Dem	US House	31	72	0.46
https://www.facebook.com/judgecarter/	Rep	US House	31	103	0.66
https://www.facebook.com/ColinAllredTX/	Dem	US House	32	819	5.22
https://www.facebook.com/petesessions/	Rep	US House	32	429	2.74
https://www.facebook.com/MarcVeasey	Dem	US House	33	52	0.33
https://www.facebook.com/billups4congress/	Rep	US House	33	227	1.45
https://www.facebook.com/UsCongressmanFilemonVela/	Dem	US House	34	31	0.20
https://www.facebook.com/profile.php?id=100011088094658&fref=mentions	Rep	US House	34	65	0.41
https://www.facebook.com/LloydDoggettTX/	Dem	US House	35	156	0.99
https://www.facebook.com/Davidsmallingforcongress/	Rep	US House	35	12	0.08
https://www.facebook.com/daynasteele36/	Dem	US House	36	227	1.45
https://www.facebook.com/RepBrianBabin/	Rep	US House	36	137	0.87

Table 4: Sample summary by page.

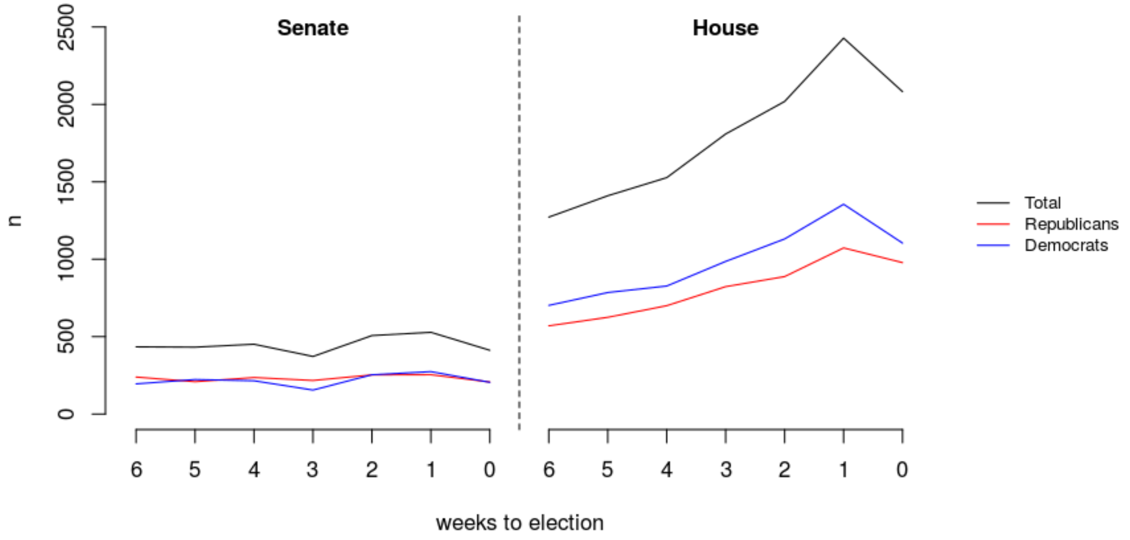

Figure 8: Sample size of digital traces collected by party and election-type, over the monitoring period. The increasing trend in collection for the house is due to increasing collection for "swing" districts, in the hope of better filling competitive cells. Weaker numbers for "election week" are due to lower number of days available for collection

B Variable Selection

Variable_Names	Class	Vote Choice Cell	P($\boldsymbol{x} = \text{NA}$)	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

Table 5: Summary table of the reduced covariate-space $\boldsymbol{X}$.

C Congressional District Plots

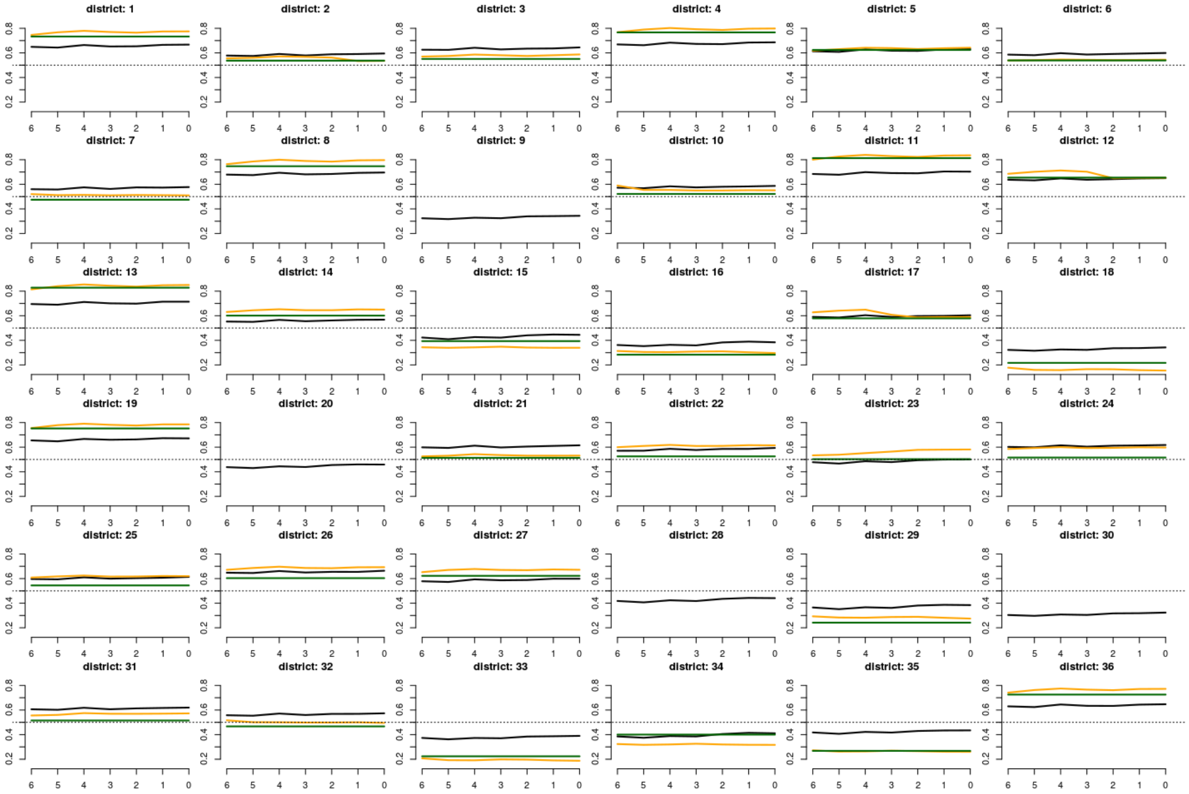

Figure 9: Congressional District Forecasts: FiveThirtyEight and Digital Vote