1

Introduction

It was 10:30am on Election Day 2012, and my numbers were telling me that President Obama might lose Ohio.

I was sitting in the Obama campaign's Analytics Boiler Room. This was the small room that set up the week before the election, only a couple of doors down from the campaign's senior leadership. It was my work space for the last few days of the campaign, tucked away from the manic energy of the other staffers, removed even from the private "Analytics Cave" where most of the data crunching had been done over the prior year.

The room was small and windowless, the kind that is normally used as an office for one or two junior level employees, maybe three if their views on personal space tended more toward the European than the American definition. But on that Tuesday, there were desks set up for 10-15 people (read: laptops), and we were all trying to figure out what was happening "on the ground." Were Obama's supporters coming out to vote at the rate we had expected? Were there signs of voter suppression? Was there any indication that the pre-election models predicting Obama's victory (both internal and external) were somehow incorrect?

As it happens, these questions were being examined using a statistical model I had built. I had come to Chicago to work for Obama six weeks earlier, and I was there specifically to determine how we would approach these questions on Election Day. Could we measure deviations from expected turnout for different groups of the electorate in real time during Election Day? If so, could we redirect our mobilization efforts to compensate and maximize our chance of winning the election?

Well before I arrived, the Obama team had been collecting and using data and "analytics" to conduct their outreach operations at a scale and level of sophistication that had never before been seen in American politics. Many of their techniques were inspired by and expanded upon earlier work, both inside (Gerber and Green, 2000) and outside (Issenberg, 2012) of academia.

But this project was different than most. Unlike, say, the problem of determining the best possible fundraising appeal, this analysis could only be done once, on the actual day of the election. Given the scale of data collection required and the level of voter participation we were attempting to measure, any "dry runs" that attempted to measure the model's performance from simulated non-Election-Day experiences were *extremely* unlikely to reflect the properties of the actual election. Moreover, there was no precedent for this type of model working properly in an earlier campaign, from which we could have borrowed proven statistical techniques or operational best practices. Determining the best way to collect the data, analyze it, and produce informative and reliable inferences was therefore an open question. In short, this was a research project. Except it was one that might have a hand in helping Barack Obama win (or worse, lose) the presidency.

And at around 10:30am, things were not looking good. Up until that point, there was not enough data for the model to point towards any strong findings. But by late morning, a mass of data from Ohio had arrived, and my model was showing that young people were *not* coming out to vote. Even worse, there were slight indications that minorities, and even Democratic primary voters, were turning out at lower rates than we had expected. While we had dismissed the early morning data, I remember one senior analyst's ominous interpretation: "It looks like this could be real." Another analyst, extremely sharp but perhaps prone to dramatic swings, memorably declared "We're f%*&ed." A senior member of the team excused himself, and I later found out that he proceeded immediately to the bathroom, in order to vomit (Alter, 2013: p. 353).

Thankfully, things turned out for the best[1]. While we were developing the model and the system, we had taken great care to not only develop the best possible estimates, but also to have a good sense of

[1] At least, if you are an Obama supporter.

the quality and uncertainty around those estimates. We were also fortunate to have outstanding senior leadership within the Analytics department, and indeed throughout the campaign. We were able to put the appropriate amount of weight on these results. Seeing these alarming trends, the campaign did not freeze, nor did it panic into changing the bulk of its long-standing mobilization plans. Instead, we diverted *some* resources while staying within the basic framework of the day's operations.

As it turns out, some of the trends diminished over the course of the day and returned to expected levels. But many of them *did* in fact come to fruition. Using data from Catalist, a firm that collects voter registration databases which include validated voting records for past elections, I can compare turnout trends from 2008 to 2012[2]. In Ohio as a whole, turnout among registered voters dropped by about 2 percentage points. Among young people aged 18-29, turnout dropped by roughly 5 points, more than doubling the state-wide decrease. Democratic primary voters also dropped by 5 points, while turnout among Republican primary voters did not drop at all. Although African Americans only dropped by about 1 point, Hispanics dropped by 4 points[3]. Despite these unfavorable turnout trends, Obama still, of course, won the state by a comfortable margin, a topic to which I will return in Chapter 4.

I open my dissertation with this story because it is reflective of a trend toward more highly quantitative research, both within academic political science and in the practical day-to-day work of political campaigns and advocacy organizations. The aforementioned research question—how to best measure deviations from expected turnout in real time on Election Day—is one that could easily be asked of political methodologists. Indeed, given my particular training in political methods and my background in practical campaigns, that is exactly the reason that I was asked to do it!

But it is also different from the bulk of the work done in political science methods as recently as a decade ago, for a number of reasons. First, it is unlikely that this problem could have been

[2] These numbers are not exactly comparable to the numbers we were examining that day, because the Obama campaign used an internal database instead of Catalist's database in 2012. But the numbers are roughly similar, as they both reflect the same basic type of data, collected from the Secretary of State in Ohio.

[3] The full numbers for all described groups: turnout dropped from 73 to 71% statewide, 56-51% for people aged 18-29, 90-85% for Democratic primary voters, 90-90% for Republican primary voters, 67-66% for African Americans, and 61-57% for Hispanics.

solved using traditional tools of public opinion research—namely, surveys—in no small part due to overreporting bias in self-reported responses about political participation. Second, and relatedly, the problem required the use of large-scale data and modern statistical methods. The system needed to interface with a modern campaign infrastructure, which these days involves databases of every registered voter in the country. Third, the question was purely operational, as opposed to substantive. We were not interested in examining causal mechanisms or explanations as to *why* turnout was high or low for different groups; we were only interested in measuring it properly and responding accordingly.

In many ways, this last piece is a critical distinction between academic political science and contemporary work in day-to-day political campaigns. When a campaign has a goal of running its operations more effectively, there are myriad ways that large-scale databases and modern statistical methods can help them remove inefficiencies and test methods of outreach, given the particularized context of that campaign. When political scientists try to learn about the fundamental nature of political behavior, it is sometimes unclear how these same data can help.

While the promise of "big data" is alluring, the truth of the matter is that a well-constructed research question and a cleanly constructed, even if "small", dataset is often sufficient, and even preferred, to achieve the goal of deriving valuable insights into broad political phenomena. As an example, consider that *The People's Choice* was written with the aid of a survey dataset of only 2,400 respondents in Erie, Ohio. This is a minuscule dataset by today's standards, but with it, Lazarsfeld, Berelson and Gaudet (1944) developed their two-step model of campaign communications, which had enormous impact. Of course, it should also be noted that there is no standard nor reasonable definition of what constitutes "big" versus "medium" versus "small" data, nor is it always clear why "big data" is so different than other traditional forms.

Even the American National Election Studies (ANES) can be thought of as "small" to some, in that it consists of a few thousand survey responses every year. Despite this, the ANES is perhaps the most important dataset used to advance our understanding of American political behavior, because (a) it has a rigorous sampling design, especially by the standards of when it began in the 1940s; (b) it has been administered in a relatively consistent fashion over the course of decades; and (c) decisions about which questions to ask and how to ask them are made with great care. All of these factors come together to make the ANES an invaluable tool, despite the relatively small number of respondents.

With that said, there *are* times when large scale data and modern statistical methods afford certain opportunities that are simply not feasible with smaller scale datasets or traditional methods, and those cases are becoming increasingly common. Indeed, the journal *Poltiical Analysis* recently published a special issue collecting articles that feature "innovative methodologies to analyze Big Data." Among them were an evaluation of Amazon.com's Mechanical Turk as a platform for experimental research (Berinsky, Huber and Lenz, 2012), the development of a Bayesian text analysis method to identify words that capture partisan differences in political speech (Monroe, Colaresi and Quinn, 2008), an investigation into the aforementioned overreporting bias in surveys on political participation using the large-scale validated voter data (Ansolabehere and Hersh, 2012), and others.

Large-scale datasets and methodologically rigorous analyses are not unique to the last decade, of course. To take one prominent example, consider NOMINATE scores (Poole and Rosenthal, 2011), which determine ideological ideal point estimates for every member of Congress throughout its history. They were first developed in 1983 and leverage a database of every single roll call vote ever recorded in both houses—a substantial dataset and well-formed method, to be sure. As another example, Erikson, MacKuen and Stimson (2002) estimate an array of structural equation models and time series analyses to develop a macro-level understanding of American politics. Even though they did not use "big data" as it might be understood today, their technical sophistication and insightful work produced a series of indicators that are influential to this day, from *macropartisanship* to *political mood*, among others. Most importantly, they were able to use these measures to provide insights into how American politics works, at the system level.

Though these examples vary by topic and details of execution, they share important common properties. They utilize relatively large-scale data, they implement advanced statistical methods which are well-suited to extract the most insight from that often messy data, and, most importantly, they use this technical sophistication to add to our understanding of political science in ways that are not plausible without the scale of the data at hand.

The three papers comprising this dissertation fall under this same broad umbrella of research. In each, I either utilize some large-scale data source, some relatively sophisticated statistical model, or both. The goal of each paper, however, is not to explore technical boundaries or to tout large datasets; the goal is to approach a classical problem in American politics and examine it through a new lens.

In the first paper, presented in Chapter 2, I bridge two often-studied topics: the "running tally" model of retrospective political evaluations, and the political socialization process. Specifically, I build a generational model of presidential voting, in which long-term partisan presidential voting preferences are formed, in large part, through a weighted running tally of retrospective presidential evaluations, where weights are determined by the age in which the evaluation was made. Under the model, the Gallup Presidential Approval Rating time series is shown to be a good approximation to the political events that inform retrospective presidential evaluations. I find that the political events of a voter's teenage and early adult years, centered around the age of 18, are enormously important in the formation of those long-term preferences.

Although this topic has been studied before, two technical advantages allow me to add to the work that has already been done. First, I aggregate more than 300,000 survey responses over 50 years. This is a substantial dataset that affords the flexibility of disaggregating responses by individual birth year (equivalently: individual age). Second, I build a rather sophisticated model connecting these survey responses to the Gallup Approval data. The model is quite flexible and captures many different aspects of the possible structure in the data. Indeed, the ability to construct such a flexible model and fit it in a reasonable amount of time has only been available in the last few years, aided by the development of powerful new software (Stan Development Team, 2013).

Importantly, though, the paper does not end with a technical description of the model and its results. The second half of the paper connects my results to a historical narrative of presidential political events from the 1940s to the present day. This is a critical piece of the analysis where we see the formation of five main generations of presidential voters. The narrative is, I hope, both familiar and clarifying—familiar by resonating with our understanding of contemporary presidential history, and clarifying by connecting that history to the new understanding of the political socialization process described by the model.

The second paper, in Chapter 3, transitions toward the use of truly large-scale data, from the perspective of political science research. I examine the classical problem of survey measurement, showing how large-scale voter registration databases, in conjunction with modern statistical methods, can substantially improve the inferences we can produce alongside political surveys. I illustrate through an example, where I use multilevel regression and poststratification (MRP) to produce vote choice estimates for the 2012 presidential election, projecting those estimates to nearly 195 million registered voters.

This paper contributes to political science methods in numerous ways. From a statistical perspective, I take a big step forward in the capabilities of MRP by increasing the breadth of the model from earlier work (Ghitza and Gelman, 2013). From a "big data" perspective, I show how inferences can be projected to a dataset of the described magnitude, and why that matters. Importantly, I also propagate uncertainty from the statistical model through the entire process, a critical piece in broadening the applicability of the process. Last, I give examples showing how the auxiliary data available through these databases can augment survey analyses.

Although Chapter 3 tackles a familiar problem—survey measurement—it is in many senses a pure methods paper as opposed to a substantive political contribution. Although I provide some examples and a discussion of how the data can be used, it is a precursor to substantive research. It can be thought of as a bridge, providing an introduction and some necessary methods that lead into the next paper, presented in Chapter 4.

Here, I examine another oft-investigated problem—how much can voter turnout levels affect partisan election outcomes? Previous work has come to mixed conclusions, and, importantly, many results may be biased due to overreporting in the survey responses that are used to measure voter turnout. I revisit the question within the context of the 2012 Presidential Election, using the inferences I derived in Chapter 3, along with additional data from the voter registration databases. These data allow me to extend the simulation framework developed by other scholars to help examine this question. In particular, (1) I use validated voter data to overcome overreporting bias; (2) the enormous sample size, along with a wide array of covariates, facilitate a detailed matching analysis that helps me investigate different plausible mechanisms for turnout change; and (3) the scale of the data allow me to examine turnout change for particular subgroups, not just for the aggregate electorate. Doing so for different demographic groups and in different states leads to important new findings as they relate to turnout change under the Electoral College. Lastly, by generalizing turnout change beyond the aggregate case, I explore more general mathematical properties that define when turnout change can impact election outcomes. This is an important contribution to our substantive understanding of the topic.

Throughout all three chapters, I employ a set of common analytical principles almost uniformly. First, when I use data to examine some relationship or trend, I use preliminary graphs to give the reader a sense of the raw data before it goes through the filter of my statistical models. Too often, in my view, papers proceed by running a simple (usually linear) model, and empirical results are derived through the interpretation of regression coefficients. This, of course, can be inappropriate when the structure of the data is non-linear! This is a simple example, but generally I have found that the most straightforward way to combat this type of problem is to actually show the data. As I am analyzing it, and as the reader is digesting my results, it is helpful to keep the structure and distribution of the raw data in mind.

Second, and relatedly, I use graphical displays of information throughout the paper to explore modeling results. These graphs are used for many purposes—to examine model results, to check whether those results fit the data, to project results to additional contexts, and so on. In my view, graphs are almost always preferred over tables when it comes to the goal of translating large amounts of data in clear and unobtrusive ways (Kastellec and Leoni, 2007).

Third, because of the scale of the data, I rarely discuss statistical significance. When I use voter registration databases, as in Chapters 3 and 4, I have 195 million observations. In this context, classical formulations of statistical significance are essentially meaningless, as there are enough observations to make almost every comparison statistically significantly different from zero. Even under traditional circumstances, statistical significance as a reliable measure for scientific discovery have come under increasing scrutiny as of late (Goodman, 2001; Ioannidis, 2005). In place of p-values and statistical significance, I tend to focus on other measures that are more appropriate to the data and the research question at hand. In all of the chapters, I focus more on substantive significance, interpreting the magnitude of the relationships in the data and how they pair with my qualitative understanding of the problem. In Chapters 2 and 3, I propagate and report uncertainty in my estimates, but again I focus more on the substantive interpretation. In Chapter 4, I conduct a series of robustness checks to test my results against different model specifications. In many cases, I compare my substantive findings to other data sources or other related studies, in order to ground my results. Although these are not hard and fast rules, I use my best judgment to present the appropriate and relevant results throughout.

Fourth, I place my results into both the context of cutting edge political science research, as well as into the realm of practical application and relevance to contemporary political practice. When building the generational model of presidential voting habits in Chapter 2, I relate it to recent evaluations of the current President, and to the implied long-term preferences of young voters who are now entering the electorate for the first time. When discussing survey measurement in Chapter 3, I do so with an eye on solving real-world problems, especially in the face of increasingly problematic survey collection and exit polls. When exploring the dynamics of turnout change and partisan electoral outcomes, I use my results to reflect on the role and limitations of political campaigns, and I discuss how effective organizations could use these findings to better allocate their resources.

Through all three papers, I hope to set up a solid foundation for future research, both for myself and for other scholars. I am particularly excited about my work using voter registration databases. Others have already contributed a valuable set of studies using this type of data in the recent past. But, in my view, this is still a largely untapped resource that could yield a range of valuable scholarly opportunities on a wide range of topics. Due to my background working in practical political campaigns, I have had the opportunity to work closely with these datasets for quite some time, gaining a unique level of access to the data. But my unique access is temporary, as the data is becoming increasingly available to academic institutions—from Catalist, for example. I hope that the work presented here can continue along the path of those who have used these data before, and lay the groundwork for it to be used more frequently in the future.

might increase rather than decrease (Lohr 2010).

This approach balances the trade off of post-stratification and raking methods, leading to more stable weights while still accounting for interactions among demographic variables. In practice, null strata are minimized, extreme weights are rare, and an analyst can easily account for many demographic variables, including four-way interactions among key variables. Additionally, analysts are afforded the flexibility of accounting for interactions among some variables, but only margins on variables that are either independent of other variables or for which only margins are known.

4.4 Application: Obama for America 2012 Analytics Polling

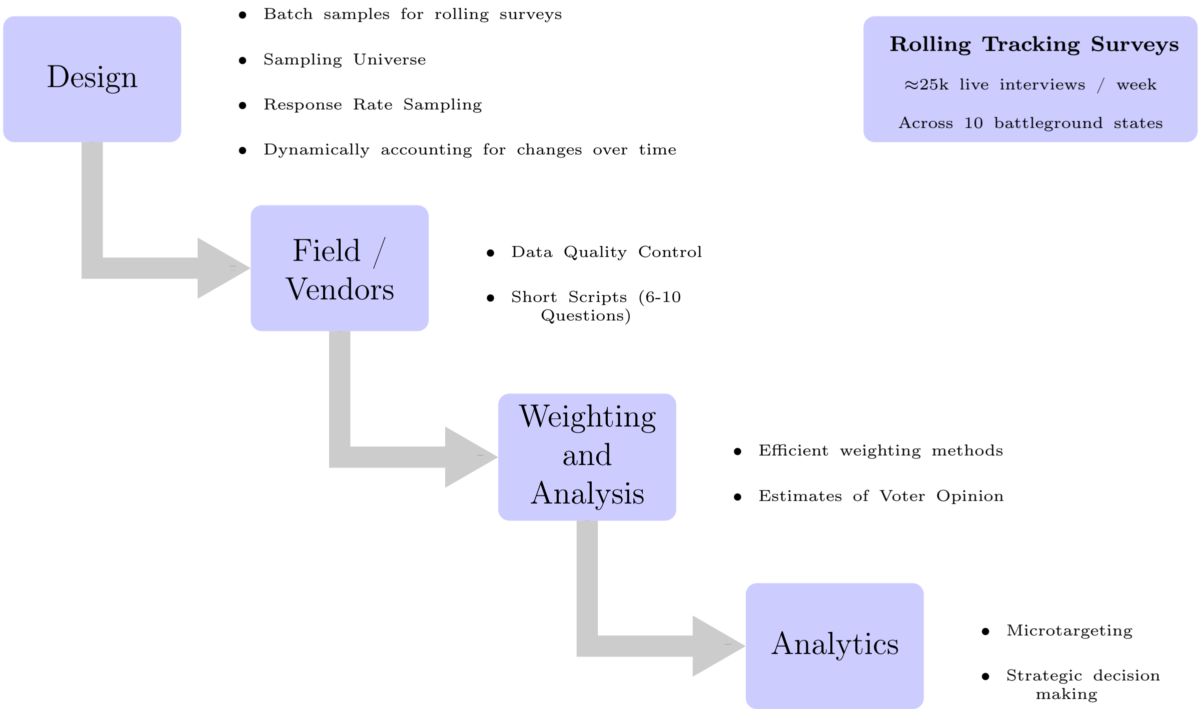
The Obama For America 2012 (OFA) campaign used a design based approach to unit non-response combined with a new flexible weighting scheme discussed in (Schickler, Berinsky and Sekhon 2012) for the large-scale polling operation fielding every day between the beginning of April 2012 and November 6, 2012. The campaign fielded live rolling, cross-sectional surveys during this time in key battleground states, ultimately contacting over one million voters through random surveys. Sampling, data collection, and weighting were conducted on a weekly basis, the process for which is outlined in Figure 4.1.

Working from an up-to-date voter file, OFA employed a list-based random sampling design of registered voters using state voter files. Random sampling of known registered voters alleviates the need for screener registration questions necessary in RDD designs, which can be subject to dishonesty and confusion of voters. Sampling was done using a stratified sampling design, including demographic variables available on the voter file. The target distribution was based on modeled estimates of the likely election day electorate.

Strata were created using seven key variables, including geographic, demographic, and behavioral measures. Stratification was done within polling regions within states. Polling regions were defined with the aim of determining areas of the state with roughly equal populations that exhibited similar patterns in voting behavior. The goal was for variance in support for the candidate to be greater across regions than within regions. Demographic measures included individual characteristics such as age, sex, race, and party registration where available. Finally, since the goal of stratification is to account both for variation in the primary outcome of interest, in this case support for the candidate, but also for behavioral measures related to self selection in to the survey, the sampling design used a proprietary proxy for political interest. This measure accounted for many factors, including previous vote history and other expressions of political engagement.

These stratification variables formed the basis of analysis for the response rate calculations, sampling design, and weighting methods. Each of the methodological steps in the polling process is outlined below.

Figure 4.1: Obama for America Analytics Polling Process


Design
• Batch samples for rolling surveys
• Sampling Universe
• Response Rate Sampling
• Dynamically accounting for changes over time
Rolling Tracking Surveys
≈25k live interviews / week
Across 10 battleground states
Field / Vendors
• Data Quality Control
• Short Scripts (6-10 Questions)
Weighting and Analysis
• Efficient weighting methods
• Estimates of Voter Opinion
Analytics
• Microtargeting
• Strategic decision making

4.4.1 OFA: Response Rate Sampling

Call data from the campaign was collected on all attempted calls, including whether or not the target voter began and successfully completed a poll. Response rates were calculated to determine the likelihood that a given voter would complete a poll if contacted. Response rates were calculated on state-specific data sets where data allowed, and among data pooled for states in similar geographic regions and with a similar data structure where state-specific data was insufficient.

Given the changing dynamics of voters' interest and engagement in the campaign, this measure of responsiveness was updated frequently. For instance, as the election neared, given the large number of calls being placed by polling firms and campaigns, many voters began experiencing call fatigue, and response rates decreased across the full spectrum of voters. There were also also events during the political season that led voters to engage at different rates. Following the first debate, Democratic leaning voters became less likely to respond to polls in the face of a poor debate performance from the President. At the same time, Republican leaning voters became more engaged, evidenced through their increased rate of response to political polls. While there may have been a small change in the support levels for the candidates, much of the change seen in public polling was driven by this differential level of engagement and enthusiasm among partisans following the debate.[4] By updating baseline levels of responsiveness, the impact of changes in enthusiasm based on campaign events were mitigated, making it easier to discern true changes in public opinion.

Response rates were calculated using Classification and Regression Trees (CART), incorporating all of the stratification variables listed in Section 4.4. The complexity parameter was cross-validated, and the minimum bucket size was set such that overfitting was minimized. While modeling response rates at an individual level is an alternative, the campaign used CART methods in order to determine the strata for which response rates were statistically different. Calculating response rates at an aggregate level like this allowed for response rates to be different for groups whose responsiveness is statistically different, yet allowed for pooling of data for smaller or highly unresponsive groups. The advantage of pooling data for highly unresponsive groups is that, when sampling inversely proportional to response rate, groups with very low response rates will be oversampled at very high rates. Having a very precise measure for non-responsive groups, where noisy estimates can lead to very high, and unstable, weights, is important to minimize costs and increase the quality of the return data.

The tree in Figure 4.2 was constructed using previous data on whether or not respondents completed a survey, and the algorithm is given five qualitative variables: political

[4] Differential rates of enthusiasm in response to may also be a sign of a changing likely electorate. Changes in enthusiasm should be incorporated in to turnout expectations. Further research needs to be conducted on how measures of enthusiasm, such as responsiveness, are correlated with likelihood to turnout.

engagement, race, age, party, and gender. Political engagement is a propriety measure, dichotomized to create a measure of high vs. low political engagement levels. This is the most predictive variable of likelihood to complete a political survey, and therefore is the first split in the tree. Within the highly politically engaged group, race is the next most predictive variable, splitting among all three races in the data set. Among the low political engagement group, African Americans and Hispanics behave similarly, so the tree only splits out Whites separately. Further splits are conducted on age, although the dividing lines vary depending on the political engagement by race classification.[5] This process continues until either there are no remaining stratification variables or the variance explained by the remaining variables is not statistically significant.

Figure 4.2: Example of Classification Tree for Response Rates to Political Surveys

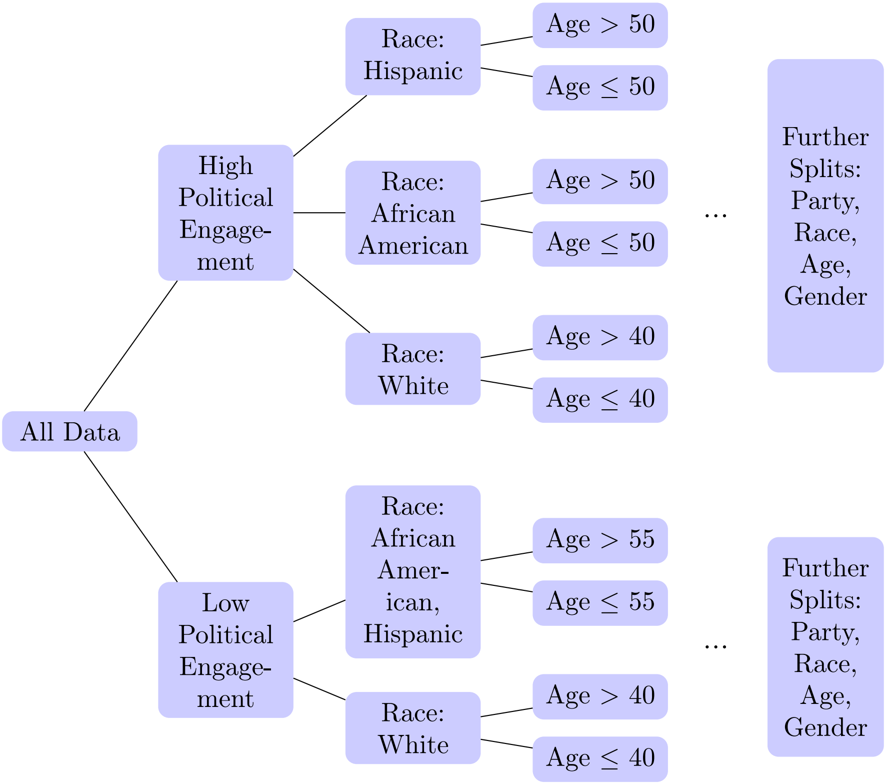


Table 4.1 shows an example of the most and least responsive groups to Obama For America's Analytics Tracking polls. There are many aspects of this table that exemplify

[5] Further splits can include additional splits on variables that have not already been fully stratified, such as additional splits on age within branches, or on strata groups that were combined earlier in the branch

Table 4.1: Examples of Response Rates from a Political Survey

Political Engagement*	Race	Age**	Party	Gender	Attempts	# Respondents	Response Rate†
Low	H	30 to 35	D, I, R	F, M, U	19,584	316	1.6%
...	...	...	...	...	...	...	...
High	W	55+	D	F	4,303	1,187	27.6 %

* Political engagement is a dichotomized variable based on a proprietary measure of engagement used by the campaign.
** Age bucket definitions have been changed slightly from the proprietary values used by the campaign
† Note that this is not conditional on contact, but rather includes the probability that the phone number is live and associated with the target respondent

typical trends in political surveys. First, the biggest predictor of likelihood to complete a political survey is captured by any measure of political engagement. Less political engaged respondents are less likely to complete political surveys. Age is also a big predictor of likelihood to complete a political survey, which is a function of both data quality and behavior. Younger respondents are more likely to be transient, and their phone numbers are more likely to be out of date. However, young respondents are also less willing to take a political survey conditional on contact. Data quality issues also give rise to differential response rates between races.

Another important thing to notice in Table 4.1 is that even though there is a lot more data in the least responsive group, the data is pooled among party and gender. Once splits have been made on political engagement, race, and age, the remaining variance is not significantly related to gender and party. The response rate is only about 1 in 62 respondents, which is very low.[6] Further splits on gender or party increase the instability in the estimates without providing any additional statistically significant information, so it is better to pool the data across strata. This is the advantage of CART methods for avoiding over-fitting the data.

The American electorate is not evenly distributed among cross classifications of these variables. For the most responsive groups, there is a lot of data among older, politically engaged, white respondents. It is important to note that response rate sampling conditions on responders, and does not address the bias due to Equation (5.2). However, in political data, there is evidence that the bias due to Equation (5.2) is small relative to the observable imbalance in Equation (5.1) (Keeter et al. 2000).

A balance must be struck between using more, older data to increase the power to detect differences in response rates and limiting to newer data that best reflects current election dynamics. In order to incorporate changes in responsiveness in response to changing election dynamics, flexibility in the measurements were allowed within terminal nodes

[6] Note that this includes disconnects and wrong numbers.

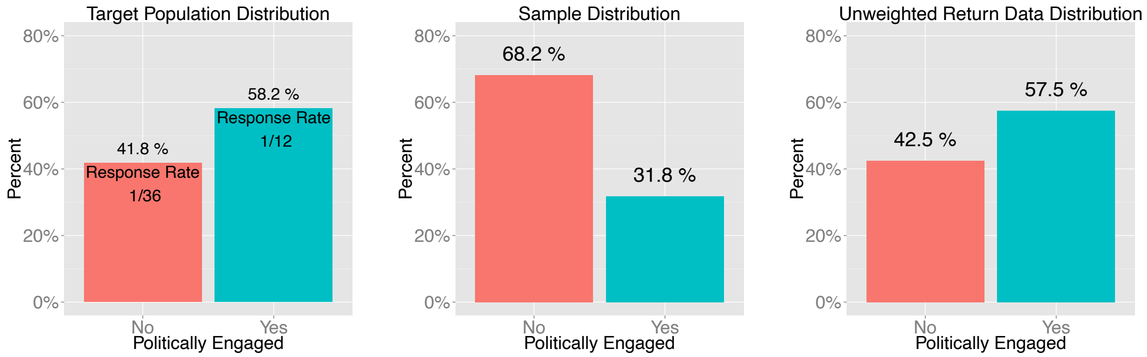
of the tree. Using the data from strata defined by terminal nodes of the CART tree, data within a node was divided into two groups based on a point of reference, such as a date three weeks prior. A t-test was conducted on data within the terminal node of the tree to determine if the response rate prior to the reference date was statistically significantly different than the response rate in the data collected after the reference date. If the response rates were significantly different, the response rate based on the most recent data was used. This allowed for flexibility in response to changing campaign dynamics, but prioritized demographic and behavioral attributes for determining response rates over the influence of the campaign cycle.

4.4.2 OFA: Checks of Design

Each week a new sample was constructed based on the refreshed response rates. OFA employed stratified sampling. Strata breakdowns were defined based on a target population, adjusted inversely proportional to estimated response rates. Since the CART method returns estimates at no smaller than a strata level, each strata defined by the stratification variables can be mapped to a terminal node in the CART tree. Within strata, a simple random sample of respondents was selected.

The final sample was randomized and calls were fielded through the week, spread equally throughout the week. Response rates varied systematically with day of the week, a factor that was not included in the campaign's sampling process but which could be included in samples for future studies. No quotas were placed on any of the stratification variables, only on the final sample size. Final data was monitored daily, and checks were conducted to ensure that the return data was representative of the target population.

Figure 4.3: Example Distributions of Sampling Process Adjusted by Response Rates Florida, 4/24/2012 - 4/30/2012



(a) Target Population Distribution with Response Rates

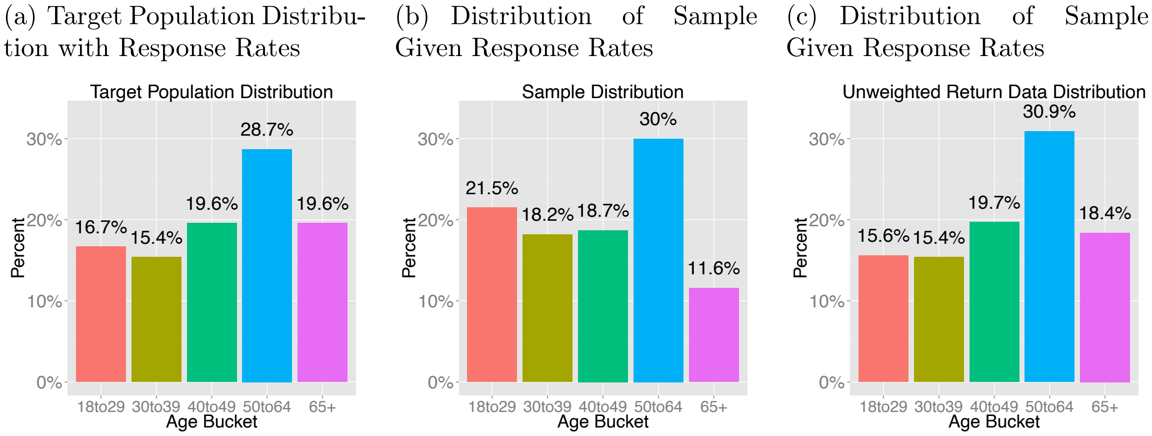
(b) Distribution of Sample Given Response Rates

(c) Distribution of Sample Given Response Rates

Response rate sampling requires that measures of non-response are accurate. Given a target population and an estimate of non-response, the final sample will not look representative of the target population, however the final, unweighted return data should be representative of the target population. Figure 4.3 shows an example of the sampling process with response rates. Figure 4.4(a) shows the target distribution of the political engagement in Florida, the state this example is drawn from. The political engagement variable is the most predictive variable in terms of explaining likelihood to complete a survey, and those who score highly on this measure are three times more likely to complete a survey than those who are not politically engaged. On average, one in twelve numbers dialed among those who classify as politically engaged will complete a survey, whereas only one in thirty six among the non-politically engaged will.[7]

Figure 4.5(b) shows the distribution of the political engagement variable in the sample. Non-politically engaged voters are oversampled relative to their size in the target population. Figure 4.5(c) shows the distribution in the unweighted return data. Despite the oversample of non-politically engaged voters in the sample, the final data is representative of the target population, with the breakdown on the political engagement variable within statistical noise of the target distribution.

Figure 4.4: Example Distributions of Sampling Process Adjusted by Response Rates Michigan, January 2012


(a) Target Population Distribution with Response Rates

(b) Distribution of Sample Given Response Rates

(c) Distribution of Sample Given Response Rates

The political engagement measure is a proxy for a behavioral driver of non-response. Response rate sampling should incorporate measures related both to behavioral predictors of non-response and demographic predictors related to the outcome of interest. Some

[7] It is important to note that the response rate includes dials to numbers that are disconnected, wrong numbers, and other issues related to the quality of the data. The response rate is not a purely behavioral measure, and conditional on a target respondent being reached, the complete rate for surveys was, on average, between one in two to one in five, depending on the state and level of engagement during the campaign.

demographic variables, such as age, are highly predictive of non-response and many outcomes of interest. Figure 4.4 shows an example of how response rates relate to age. Young voters are more unresponsive than older voters, with those over the age of 65 significantly more likely to respond to a political poll than the younger cohorts.

As seen in Figure 4.3, Figure 4.4 shows that, despite a large differential in the likelihood of different age cohorts to respond to an OFA survey, adjusting for this known difference in the sample using response rates allows for the raw return data to be highly representative of the target population. Response rates account for both behavioral and quality of data issues that impact responsiveness to surveys. While political engagement factors are driven primarily by behavioral predictors of responsiveness, factors such as age are confounded by behavioral and data quality factors. Young voters are more likely to have old phone numbers associated with their names on the voter file, such as the phone number of their childhood home if they register at 18. More transient populations are more likely to have out-of-date contact information, and factors related to factors such as transience will confound behavioral and data quality factors.

4.4.3 OFA: Weighting

Sections 4.4.1 and 4.4.2 outlined how response rate sampling can increase the representativeness of survey return data. However, due to changes in voters' engagement during the election, changing dynamics of underlying responsiveness, or statistical noise, there are still small imbalances in the raw return data. Section 4.3 outlines various methods for *post hoc* weighting adjustments. Weighting return data can improve the accuracy of surveys, although extreme weights that occur in highly unrepresentative samples can induce noise in estimates.

OFA employed raking for *post hoc* weighting. Weighting was conducted using the same variables used in the the sampling frame. As shown in section 4.4.2, the return data was highly representative due to the response rates accounted for in sampling. This, combined with a flexible weighting method, allowed for an increased number of variables to be accounted for in the weighting step.

By using a raking method that accounted for all three-way interactions of all key variables, including demographics such as age, gender, race, political engagement, and region, the weighting method could account for remaining discrepancies between the raw return data and the target population. The representativeness of the raw return from week to week meant that the same factors could be accounted, minimizing the impact of an analyst's decisions on factors to control for on the estimates of public opinion.

Additionally, despite including numerous variables in the weighting adjustments, the flexible weighting method and representative samples meant that the final weights did not suffer from extreme weights. Typically, across states and time, 95% of weights were below 1.5, meaning the design effect of weighting on the survey was small and the noise in the estimates induced by the weighting step was minimal.

Figure 4.5: Distribution of Weights in Final OFA Surveys

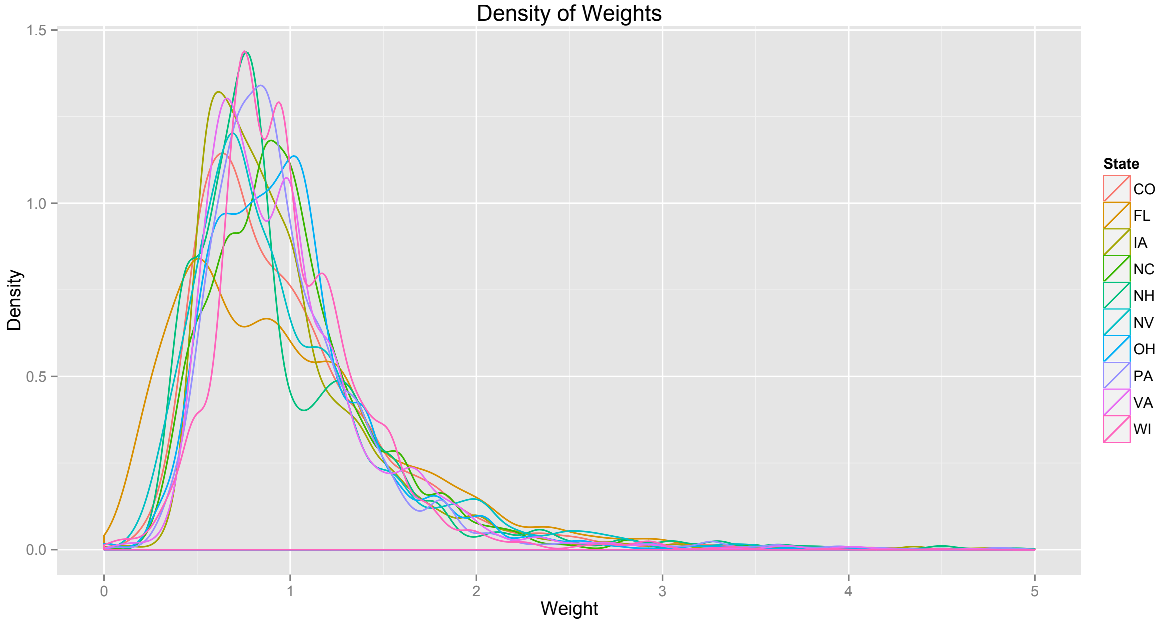


Figure 4.5 shows the density of weights by state for the final polls released by OFA. These polls were conducted between October 28, 2012 and November 4, 2012. What is important to note is that no weights exceed 5, and the bulk of the density is near 1. Most of the adjustments to the data were minor, despite accounting for up to four way interactions of key demographic variables.

The combination of response rate sampling and a flexible weighting method allowed for the campaign to have consistent, representative surveys that minimized the impact of non-response on the estimates. Estimates of public opinion remained steady over the course of the campaign. Changes that were observed were more easily attributed to changes in public opinion instead of changes in underlying demographics or differential voter engagement. This resulted in highly accurate estimates of public opinion, allowing the campaign to efficiently allocate resources. Ultimately, the polling proved to be not only consistent, but highly accurate.

4.5 Discussion and Conclusion

This chapter outlined multiple design based approaches to minimize bias due to unit non-response, as well as numerous weighting methods to account for observable imbalance in final survey returns. A new design based approach, response rate sampling, is outlined. This method uses response data from previous surveys to estimate the the likelihood a strata or individual will complete a survey. Sampling can then be done inversely proportional to these known response rates, increasing the likelihood the survey response data will be representative. This minimizes the impact of *post hoc* weighting methods, as well as reduces the number of impactful substantive decisions a practitioner must make on how to conduct *post hoc* weighting. Additionally, in the face of observable imbalance, a new approach to raking methods is discussed.

While this chapter addressed the advantages response rate sampling serves in reducing bias due to unit non-response, especially with respect to the bias outlined in Equation (5.1), there is further work to be done on the impact response rate sampling has on variance reduction. Response rate sampling increases the representativeness of the final data, decreasing the emphasis placed on *post hoc* weighting adjustments. This reduces variance, and minimizes the likelihood of extreme weights. However, response rate sampling may also lead to increased confidence in the assumption that data is missing at random (MAR). The MAR assumption states that nonresponse depends only on observed variables (Lohr 2010), since hard to reach groups are overrepresented in the sample, increasing the likelihood that these groups are not represented purely by unrepresentative early responders who are given high weights. Further work will explore the impact of response rate sampling on this assumption, and its impact on variance reduction of estimates of opinion.

In addition to introducing response rate sampling, this chapter applied this method in the context of a large scale presidential campaign. Obama for America's (OFA) 2012 internal polling employed response rate sampling in the design of their internal analytics polling operation. The campaign spoke with over one million voters across battleground states, collecting information on calls placed to over ten million voters in the context of random surveys. This massive polling operation allowed for frequent refreshes of the CART models estimating the probability of response among demographic strata. By incorporating up-to-date estimates of responsiveness in the sampling design, OFA was able to minimize the impact large campaign events, such as the first presidential debate, had on estimates of public opinion. Large events such as these impact not only voter opinion, but voter's enthusiasm, which translates into differential rates of response among different partisan and demographic groups. Using accurate baseline measures of responsiveness, updated in the face of voter fatigue towards surveys that increased over the course of the campaign, OFA was able to mitigate the changes due to unit non-response in internal surveys, and keep a consistent and accurate measure of public opinion.

Response rate sampling also allowed the campaign to collect representative polling data, minimizing the impact of *post hoc* weighting methods. Weighting methods often assign high, unstable weights, in political polls, to groups that are heterogenous in their opinions, such as younger voters and hispanics. By focusing on collecting observably representative survey data, the weighting methods were both able to account for more demographic characteristics while also keeping the weights stable.

By incorporating previous data, response rate sampling, and a novel approach to raking weighting methods, the OFA analytics polls allowed the campaign to have stable estimates of voter opinion, even in times that public polls showed wild swings. Combined with careful analytical approaches to a consistent and accurate target election day electorate, these methods allowed the campaign to have highly accurate measures of public opinion.

TO KNOCK OR NOT TO KNOCK

Experiment-Informed Voter Contact (https://web.archive.org/web/20160403142353/http://0ptimus.com/category/experiment-informed-voter-contact/)

21/12/15

To Knock Or Not To Knock

ampaigns should do to win. Often times, this coverage takes the form of obse the machinations of a particular campaign, comparing it to a study or set of studies, and then asking the question "Why aren't these guys doing what the study says gets you more votes?"

For example, many reporters have recently discovered the work of political scientists who are quantifying the effect of campaign tactics. This branch of academia, started by Don Green and Alan Gerber, has shed light on what works and what doesn't work in campaigns. At Øptimus, the firm I work at, the books and papers emanating from this branch of political science are required reading. To put it lightly, we are big fans. That being said, you have to be careful in how you interpret and apply the findings from this branch of work.

Take for example, what the literature tells us about the efficacy of volunteer operations. We often ask volunteers to knock on doors and deliver messages that increase the likelihood that a voter will turn out to vote. Academic research has repeatedly demonstrated that competently run door knock operations can generate a measurable increase in turnout. At my firm, we have put this theory to the test in live-fire targeted races, and verified this finding. To be specific, we generally observe that you can expect about a 2-4% lift-in-turnout over a control group (a group you did nothing to) from volunteer door knocking operations. To translate that into English, if you knock on 1000 doors and attempt to have good-quality interactions with voters, you will get 20-40 people to turn out who otherwise would not have.

So what do we take away from this? Some who cover campaigns see this as a solution to winning campaigns. They see the simple and obvious prescription for any campaign to execute if it wants to win. They say door knocks need field offices, and so if we look at who has field offices, we know who is running a smart campaign, and who isn't. The problem is, this is declaring a cookie-cutter solution to winning campaigns that fails to acknowledge a lot of complexity that a campaign must face. In ways, this is no better than if one were to read about positive effects of fish oil on heart health, self-medicate without speaking with doctors, and then never understand that while fish oil is indeed helpful, exercise, diet, and statins are pretty important too.

Let's consider a hypothetical case study. First, let's accept as a given you can get a 4% increase in turnout on those you door knock if you can place that door knock within 7 days of the election date. Now let's start bringing that assumption into a real-world situation a campaign might face. Consider the imaginary case of a contested primary race in a midwest state with a field of 10 candidates.

lection) is 150,000 people. In this hypothetical case, let's assume the front ru as 30% of the vote. Further, let's assume you work for a candidate with 20% of the vote. What this means is that you have 30,000 supporters within the total universe of people who will vote. While you know that these 30,000 must exist, you don't know exactly who they are. So you do lots of phone calls to ask people who they will vote for (in campaign terms, this is known as "voter ID calls"). You make calls for months and attempt to contact everyone to find your 30,000, but at the end of the day you discover that some people just aren't reachable. This results in you being able to find and confirm 20,000 supporters.

Now, let's say you run some tests and find out that among your base of supporters, you can get 2% of them to do volunteer work at your campaign offices. This means inside your supporter base of 20,000, you've got 400 folks who could work for you as volunteer at some point in the campaign. Let's also say from past races that you have learned that on average you can get a volunteer to perform a 3-hour shift of door knocking in the 7 day period leading up to an election (some people will do 5 shifts, many people will never show up for a shift in the final 7 day period, but average will be about 1). Let's say you also run the math on the geography of this electorate living in this very rural state, and surmise that on average you can hit 30 doors with one volunteer hour (including driving to door knock site, walking between houses, taking coffee breaks, etc.).

If you crunch all the numbers, you find out you've got 1200 man hours, which can produce 36,000 door knocks. You start to make door-walk lists, only to realize that since this is a very rural state, 2,000 of the 20,000 supporters you have identified are not going to be reachable. So you've got 18,000 voters you can actually reach, and the ability to hit 36,000 doors. So you try and hit everyone twice. Now because you hit everyone twice, let's say you achieve a 5% increase in turnout (instead of the 4% we might expect with only one attempt). For your work, you've generated 900 votes (5% times 18,000). Let's assume that this state actually turns out 142,000 voters in its election. This 900 votes you generated means you got an additional 0.6% (six tenths of a percent) of the vote for your candidate.

To be sure, 0.6% of the vote is nothing to sneeze at, and many elections are won by far less. The problem is you've missed the elephant in the room. Your candidate only has 20% of the vote, where the front runner has 30%. You've got a 10 point gap you need to close. While getting 0.6% of the vote produced via extra turnout is important, you are not even in the game until you have closed the 10% gap that must be closed via persuasion.

012. You know the race in a virtually 50/50 country is going to be very tight. F[illegible] let's say you can recruit 5% of your supporters to be volunteers, and on average they will work 2 shifts in the final 10 days. Further, let's assume you can ID 80% of your supporters over 9 months leading into the election. Further, let's assume that you can knock on 60 doors per hour, because your voters don't live in rural farmlands but in dense cities. Suddenly, the productivity of door knocking changes in a drastic and favorable way, and the produced results are projected to be of a scale that is critical to the strategic problem the campaign is facing (winning a 50/50 election by a hair).

All this is not to say that door knocking doesn't work. We know it does. But it like every other tactic must only be applied in support of a strategy designed to the realities that each unique campaign must face.

SHARE THIS POST

(HTTPS://[illegible]0PTIMUS.COM/TO-KNOCK-OR-NOT-TO-KNOCK/)

LEAVE A COMMENT

Your email address will not be published. Required fields are marked *

Comment

* Name

* Email

Post Comment