Why Political Polling is not Dead – a Plea for Non-Probability Polling, Algorithms, and Big Data

Tobias B. Konitzer and David M. Rothschild

1 Introduction

Not long ago, Zukin (2015) asked in the New York times: "What is the matter with political polling"? The question seems justified. Traditional political polling, as of 2016, had just endured a number of spectacular embarrassments around the world. Exhibit A: Benjamin Netanyahu won the Israeli elections comfortably, despite trailing to the Zionist Union alliance for weeks in the polls. Exhibit B: David Cameron and his Conservative Party formed the first Conservative majority government since 1992 after gaining the majority of the vote in the 2015 parliamentary election – in the face of opinion polls that predicted a razor-thin election outcome. Exhibit C: In 2016, Great Britain stunned the world by voting in favor of the Brexit – a referendum in favor of leaving the European Union – despite opinion polls suggesting that the most likely outcome was to remain in the Union.[1] Opinion polls have been of central importance for political reporting, especially during political campaigns when horse-race coverage is dominating cycle after cycle (Patterson, 1993). In turn, the crisis of polling has inevitably transpired to be a crisis of political journalism as well – at least when reliant on political polls.[2] But the relationship between polls and journalists is a symbiotic one: Polling firms need publicity, and journalists need numbers to attract readers. Hence, it is unlikely that political polls will disappear from the headlines of political reporting.

Instead of simply warning against the dangers of polls or urging journalists to change their model of political reporting, we demonstrate the reason behind the crisis of political polls. We then propose a remedy: So called non-probability polling, in combination with algorithms and Big Data. We showcases the usefulness of such a Big Data approach and demonstrate advantages when it comes to a) cost, b) time and c) fine-grained estimates of public opinion. In the applied part of the paper, we demonstrate the usefulness of non-probability polls collected via means as disparate as gaming systems, opt-in online polls, and cell phone polls. In concluding, we discuss ways to streamline this approach for political reporting.

2 Traditional Political (Probability) Polling

Public opinion polling today revolves around the notion of representative sampling, the idea that every individual in a particular target population, such as registered or likely US voters, should have the same, non-zero probability of being included in the poll. From address-based sampling

[1] http://www.businessinsider.com/yougovs-polling-methods-were-flawed-2016-5

[2] Not surprisingly, official journalistic guidelines urge journalists to treat polls with caution, see http://journalistsresource.org/tip-sheets/reporting/polling-fundamentals-journalists

in the 1930s to Random Digit Dialing (RDD) in the mass media era, polling companies have put immense efforts into obtaining representative samples (Wang et al., 2015). The adoption of RDD polling by both, polling companies and respectable media outlets, can be traced largely back to a flat-out polling disaster: In the 1936 US presidential election, the magazine *Literary Digest* predicted a Republican landslide based on two Million mail-in surveys the magazine had sent out to its readers, but Roosevelt got reelected comfortably instead.

What went wrong? Unsurprisingly, the *Digest's* pool of respondents was highly biased: it consisted mostly of auto and telephone owners, a demographic skewing wealthy and Republican. In effect, the magazine had predicted the vote choice of a a hypothetical Republican population, and unsurprisingly found that it favored a Republican president. In the wake of the 1936 elections, non-representative or non-probability polls quickly gave way to considerably smaller but representative polls. But the gold standard of probability polling has suffered in recent years. First, the most popular form of probability-polling, RDD, endured a radical decline in response-rates.[3] Low response rates are not an issue per se, but if non-response is related to political beliefs, the sample becomes dangerously biased, even if the initial sampling frame is representative. What's more, probability polling is increasingly expensive, and logistically difficult to field. These trends in combination with off-the-shelf capabilities to handle Big Data, call for a new assessment of convenience polls, 80 years after the *Literary Digest* debacle.

3 Unorthodox Political (Non-Probability) Polling

The idea to derive representative estimates from convenience samples seems ludicrous at first, but can be done based on a simple statistical assumption. The technique, known as model-based-post-stratification (MP), rests on the belief that people's political opinions can be exhaustively predicted by a vector of background characteristics, such as party identification, race or education. If one is able to attract a sufficient sample, convenience samples allow for obtaining representative estimates thanks to MP. Let's focus on a concrete example: vote-intention. MP obtains probability estimates of voting Republican or Democrat for every desired demographic cell, as specified in Equation 1 (e.g. black college-educated females living in the South, who affiliate with the Republican Party).

$$
\begin{aligned}
\mu_i &= \beta^{sex} * sex_i + \beta^{race}_{race(i)} + \beta^{educ}_{educ(i)} + \beta^{party}_{party(i)} \\
\beta^{dma_k} &\sim Normal(\beta^{vote_prev}_{dma(k)} + \beta^{Division}_{dma(k)}, \sigma^2_{dma}) \\
y_i &\sim Bernouli(g(\beta_0 + \beta^{dma}_{dma(i)} + \mu_i)) \\
\beta &\sim Normal(0, \sigma_\beta) \\
\sigma_\beta &\sim Cauchy(0, 5)
\end{aligned}
\tag{1}
$$

The strength of this multi-level Bayesian model is that it can pool statistical power. For instance, in the example above we use the overall likelihood of voting Republican by blacks to partially inform our sub-group estimates. As a second step, we can weight these group-level probabilities based on the known proportion of that demographic in the desired target population. "Known" is a loaded word, but in most scenarios we can get relatively close. Ideally, we simply take the known proportion from so called voter files,

[3] While some researchers have estimated response rates at about 36% as late as 1997, single-digit response rates are now more common than exceptional, http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/

detailed background information on every registered voter in the US. In the remainder of the study, we use voter file data from TargetSmart, a commercial data vendor providing information about every registered voter (N = 146,311,000) in the United States, supplemented with commercial data, such as party registration and a number of demographic variables. Unfortunately, it is unrealistic for journalistic outlets but those with abundant financial resources to get access to such proprietary data. But, as we show below, individual-level data from the American Community Survey (ACS) – data that is in the public domain – does as well and in some cases even better in combination with convenience samples. In Figure 1, we display post-stratified estimates to known civic benchmarks, collected via 10 different vendors of convenience samples for a number of items – ranging from whether one voted locally to whether one is active in one's community – relying on the ACS and the voter files. The results (expressed as difference from the truth) show that 1) MP can eradicate the selection bias associated with convenience samples, and 2) ACS, data that is readily available in the public domain, is apt for deriving the target population(Messing et al., 2016).

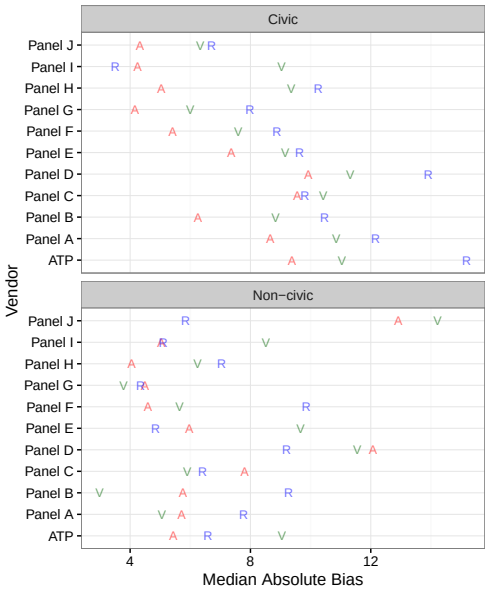

Figure 1: Post-stratified estimates ACS vs. Voter File

4 Three Examples of Useful Convenience Samples

As noted before, MP is a useful approach for obtaining small-area estimates. For example, given MP, we can estimate Republican vote intention among black college-educated females living in the South, who affiliate with the Republican Party, down to the census block level. For obvious reasons, probability polls with their limited sample sizes are not able to derive comparable small-area estimates. Below, we showcase the use of such small-area estimation, using a convenience sample collected on the Xbox gaming system during the 2012 election (Gelman et al., 2016). In short, during the 2012 U.S. presidential campaign, we conducted 750,148 interviews with 345,858 unique respondents on the Xbox gaming platform during the 45 days preceding the election. Xbox Live subscribers who opted in provided baseline information about themselves in a registration survey, including demographics, party identification, and ideological self-placement. According to some estimates, the poll came close to a 3%-median error for state-level predictions (Wang et al., 2015). Figure 2 displays the Obama-over-Romney margin in vote intention at the media market level. The over-time variation could easily be tied to ad spending, again data that is (more or less) readily available in the public domain, and would prove very fruitful for any political coverage

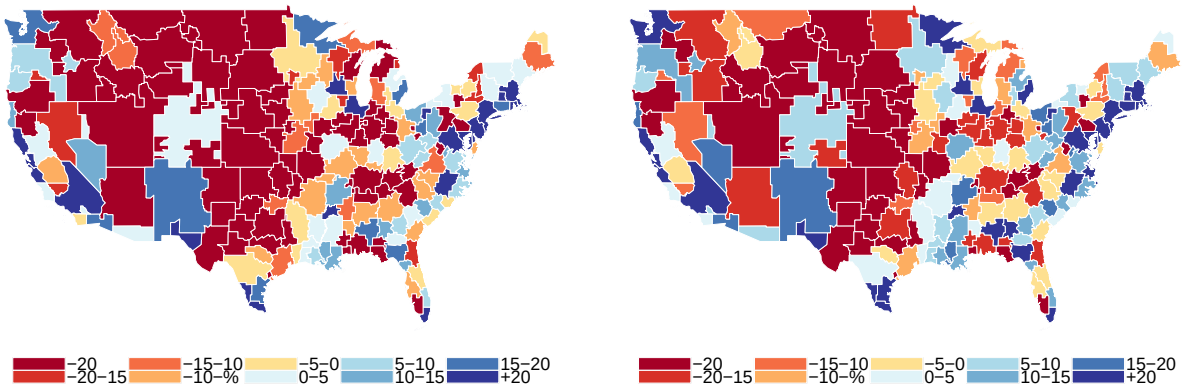

Figure 2: Obama Margin in Vote Intention over Romney for 09-22 (left), 10-22 (right) in percentage points

As a second example, we use a poll collected on MSN.com in June 2016 to showcase how day-by-day state-level variation in vote intention can be derived with relatively little effort. We asked the standard headline question that will dominate our political lives for the next five months: Who will you vote for president? And, over 1 Million MSN readers (!) answered the poll within a day. The raw data indicates that 31% would vote for Hillary Clinton, while 49% would favor Trump – a bias that is predictable given that the sample skews older, whiter, and more Republican. However, as we show in Figure 3, MP produces meaningful by-state variation, and is able to correct the heavy selection bias.

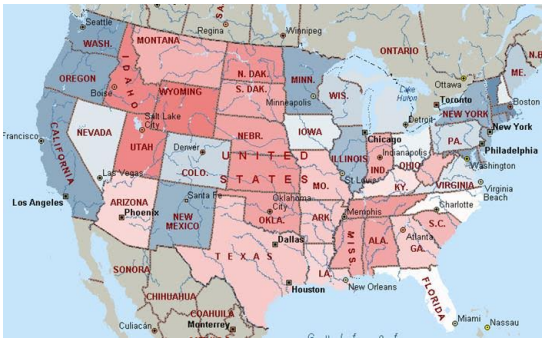

Figure 3: Post-stratified by-state estimates of vote intention from the MSN poll

One last example should serve to highlight my point. We collected vote intention for the 2016 presidential election on Pollfish, a start-up mobile survey platform. Pollfish's survey platform delivers online surveys to almost 200 million smartphone owners, or potential respondents. We specifically target smartphone owners in the United States. We show swings in vote intention at the national level in Figure 4, compared to the HuffPollster average. The most recent data suggests Clinton in the lead by 6 percentage points. [4] Importantly, cell-phone surveys also allow us to track the whereabouts of respondents very precisely. The left panel of Figure 4 showcases the whereabouts of one individual in our panel. It is easy to see how individuals can be located down to street blocks, enabling targeted questioning. For example, one could poll respondents in a given community on matters of community policy, with relatively little effort.

[4] http://elections.huffingtonpost.com/pollster

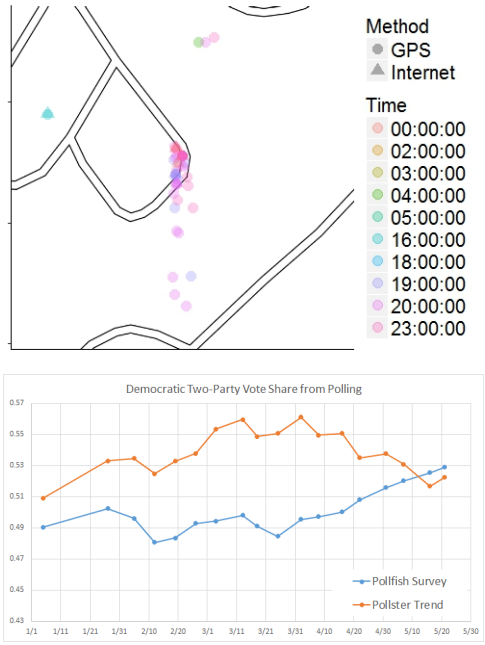

Figure 4: Geographical targeting via cell-phone polling (top), Measuring vote-intentions via cell-phone polling (bottom)

5 Conclusion

We have shown that non-probability, or convenience samples, have the potential to be viable estimates of public opinion. Of course, many predictions made here have to be validated against the outcomes in November 2016, yet the validations presented in Figure 1 and elsewhere prove promising (Gelman et al., 2016). In addition, convenience samples have the advantage of being cheap and quick to turn around. For example, given the access to high-powered computing systems and advanced statistical tools – Stan ("Sampling Through Adaptive Neighborhoods"), the wrapper used to fit results here, fits Bayesian models fast in relying on Hamiltonian Monte Carlo (HMC) implemented via the No-U-Turn (NUTS) algorithm in C++ – even the MSN data can be turned around in a matter of hours, from launch to presentation of results. Are convenience polls *better* than RDD-based polls? This is, in some ways, the wrong question to ask. It is a fair statement that given appropriate analytics, access to Big Data and algorithms, convenience polls are certainly not worse. What's more, convenience polls allow for targeted question, dynamic tracking of public opinion at very granular levels, and are cheap, both when it comes to time and effort and money. In consequence, convenience polls, as discussed here, have the potential to expand political reporting significantly.

References

Gelman, Andrew, Sharad Goel, Douglas Rivers and David Rothschild. 2016. "The Mythical Swing Voter." *Quarterly Journal of Political Science* 11(1):103–130.

Messing, Solomon, Tobias Konitzer, Andrew Mercer, Courtney Kennedy and David Rothschild. 2016. "Online Polling – Cost, Speed, Accuracy and Open Access." Presentation at the American Association for Public Opinion Research (AAPOR).

Patterson, Thomas E. 1993. *Out of order*. A. Knopf.

Wang, Wei, David Rothschild, Sharad Goel and Andrew Gelman. 2015. "Forecasting elections with non-representative polls." *International Journal of Forecasting* 31(3):980 – 991.

Zukin, Cliff. 2015. "Whats the Matter With Polling?" *The New York Times* .