



Contents lists available at ScienceDirect

## International Journal of Forecasting

journal homepage: [www.elsevier.com/locate/ijforecast](http://www.elsevier.com/locate/ijforecast)

## Forecasting elections with non-representative polls

Wei Wang<sup>a,\*</sup>, David Rothschild<sup>b</sup>, Sharad Goel<sup>b</sup>, Andrew Gelman<sup>a,c</sup><sup>a</sup> Department of Statistics, Columbia University, New York, NY, USA<sup>b</sup> Microsoft Research, New York, NY, USA<sup>c</sup> Department of Political Science, Columbia University, New York, NY, USA

## ARTICLE INFO

## Keywords:

Non-representative polling  
 Multilevel regression and poststratification  
 Election forecasting

## ABSTRACT

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

© 2014 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

## 1. Introduction

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

The wide-scale adoption of representative polling can be traced largely back to a pivotal polling mishap in the 1936 US presidential election campaign. During that campaign, the popular magazine *Literary Digest* conducted a mail-in survey that attracted over two million responses, a huge sample even by modern standards. However, the magazine incorrectly predicted a landslide victory for Republican candidate Alf Landon over the incumbent Franklin Roosevelt. In actual fact, Roosevelt won the election decisively, carrying every state except for Maine and Vermont. As pollsters and academics have pointed out since, the magazine's pool of respondents was highly biased: it consisted mostly of auto and telephone owners, as well as the magazine's own subscribers, which underrepresented Roosevelt's core constituencies (Squire, 1988). During that same campaign, various pioneering

\* Corresponding author.

E-mail addresses: [ww2243@columbia.edu](mailto:ww2243@columbia.edu) (W. Wang), [davidmr@microsoft.com](mailto:davidmr@microsoft.com) (D. Rothschild), [sharadg@microsoft.com](mailto:sharadg@microsoft.com) (S. Goel), [gelman@stat.columbia.edu](mailto:gelman@stat.columbia.edu) (A. Gelman).

<http://dx.doi.org/10.1016/j.ijforecast.2014.06.001>

0169-2070/© 2014 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

pollsters, including George Gallup, Archibald Crossley, and Elmo Roper, used considerably smaller but representative samples, and predicted the election outcome with a reasonable level of accuracy (Gosnell, 1937). Accordingly, non-representative or “convenience sampling” rapidly fell out of favor with polling experts.

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

In this paper, we show that, with proper statistical adjustments, non-representative polls are able to yield accurate presidential election forecasts, on par with those based on traditional representative polls. We proceed as follows. Section 2 describes the election survey that we conducted on the Xbox gaming platform during the 45 days leading up to the 2012 US presidential race. Our Xbox sample is highly biased in two key demographic dimensions, gender and age, and, accordingly, the raw responses disagree with the actual outcomes. The statistical techniques we use to adjust the raw estimates are introduced in two stages. In Section 3, we construct daily estimates of voter intent via multilevel regression and poststratification (MRP). The central idea of MRP is to partition the data into thousands of demographic cells, estimate voter intent at the cell level using a multilevel regression model, and finally aggregate the cell-level estimates in accordance with the target population’s demographic composition. One recent study suggested that non-probability samples provide worse estimates than probability samples (Yeager et al., 2011), but that study used simple adjustment techniques, not MRP. Even after getting good daily estimates of voter intent, however, more needs to be done to translate these into election-day forecasts. Section 4 therefore describes

how to transform voter intent into projections of vote share and electoral votes. We conclude in Section 5 by discussing the potential for non-representative polling in other domains.

## 2. Xbox data

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

Despite the large sample size, the pool of Xbox respondents is far from being representative of the voting population. Fig. 1 compares the demographic composition of the Xbox participants to that of the general electorate, as estimated via the 2012 national exit poll.<sup>1</sup> The most striking differences are for age and sex. As one might expect, young men dominate the Xbox population: 18- to 29-year-olds comprise 65% of the Xbox dataset, compared to 19% in the exit poll; and men make up 93% of the Xbox sample but only 47% of the electorate. Political scientists have long observed that both age and sex are strongly correlated with voting preferences (Kaufmann & Petrocik, 1999), and indeed these discrepancies are apparent in the unadjusted time series of Xbox voter intent shown in Fig. 2. In contrast to estimates based on traditional, representative polls (indicated by the dotted blue line in Fig. 2), the uncorrected Xbox sample suggests a landslide victory for Mitt Romney, reminiscent of the infamous *Literary Digest* error.

## 3. Estimating voter intent with multilevel regression and poststratification

### 3.1. Multilevel regression and poststratification

To transform the raw Xbox data into accurate estimates of voter intent in the general electorate, we make use of the

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

Download English Version:

<https://daneshyari.com/en/article/7408429>

Download Persian Version:

<https://daneshyari.com/article/7408429>

[Daneshyari.com](https://daneshyari.com)