Using Smartphone Polling to Forecast the German Federal Election 2017*

Comparing the performance of digital and traditional polls

Moritz Hemmerlein & Alexander Sacharow

12 May, 2017

Abstract

Using online surveys to forecast election outcomes imposes severe challenges to pollsters. Non-representative samples and likely voter bias skew gathered information and require adequate statistical adjustment. This paper compares different weighting methods to obtain a forecast from raw data, collected through smartphone polling and uses the average of leading polling institutes' polls as a benchmark. We find that post-stratification including revealed last votes are an effective tool to adjust polls while post-stratification purely on demographic factors cannot offset the sampling bias for polls with 1000-2000 responses.

Contents

*The R scripts and large parts of the raw data are available on our GitHub repository. For the remaining raw data please contact the authors, corresponding address: m.hemmerlein@mpp.hertie-school.org.

1 Introduction

The digitalisation is challenging the way polling was done for many decades. Calling people on their landline phones has become difficult as response rates dropped and number of households equipped with landline phones is decreasing (Skibba 2016). In response, pollsters have resorted to other polling-methods ranging from face-to-face interviews to mobile phone calling. However, these methods either face similar difficulties to ensure representativeness or are too expensive for regular polling. In order to tackle these obstacles, pollsters are increasingly using online polls, which are cheap and fast, but can be highly non-representative (e.g. W. Wang et al. 2015).

The eventual aim of pollsters in online as well as traditional polling is to collect sample data that reflects the view of a population of interest. The major difference between both methods is that for various reasons online polls cannot ensure representativeness before the actual poll takes place. For instance, respondents of online surveys are more likely to be from certain demographic groups or share a particular political background depending on the website or app where the survey is conducted. However, such non-representative polls can be statistically adjusted to match the demographic composition of the population.

Additionally, online election polls, like traditional ones, face another problem. Election forecasters are naturally not only interested in the population as such, but in the population of actual voters. By the time a poll is made representative in demographic terms, it is still in question whether it reflects the group of people who actually cast their ballots. This, however, is crucial in order to make an accurate prediction. Traditional polling tries to account for this using likely voter models and could perform fairly well (Gallup 2010, Keeter, Igielnik, and Weisel (2016)). Online surveys will also have to be adjusted to actual voting population in order to provide accurate predictions.

In this paper, we employed different weighting strategies to an online poll, exploiting respondents usage of smartphone apps. We evaluate our methods against the rolling average of polls from leading German polling institutes. While the raw data of the online survey performed fairly poor compared to our benchmark, we show that with proper weighting we could obtain results that are close to the major polls. Moreover, we compared a rather direct and a indirect method. First, we used information on self-reported vote decision at the German federal election 2013 and data from official exit polls to construct our weights. In a second, more indirect approach, we constructed weights to adjust our demographic clusters to the German census and subsequently identified likely voters. As our results show, the direct method performed considerably better than the indirect weighting through census data.

The structure of the paper is as follows: Section two will survey the literature on non-representative polling and presents approaches to employ such data to forecast elections. Subsequently, in chapter three, we present our data and discuss idiosyncratic problems and potential sources of bias. Section four shows the methods we applied to our data. Chapter five is presenting and dicussing our results and their performance against the benchmark. Chapter six concludes and gives an outlook on practical obstacles and methodological issues that our approach suffers from.

2 Related Literature

Traditional polling and in particular election polling has relied heavily on telephone surveys for the last decades. To ensure representativeness, the standard was randomized digit dialing (RDD). The selection of random respondents was intended to eliminate the sample bias of the survey. However, for several reasons this approach has become unreliable. First, response rates have declined heavily (Keeter et al. 2006; Holbrook, Krosnick, and Pfent 2007). The Pew Research Center (2012) reported that in the U.S. response rates dropped down to 9% in 2012, compared to 36% in 1997. This is fostered by technical changes that make it possible to identify the caller before taking the call and increased the likelihood of people not answering survey requests. Second, more and more people do not get landline telephones after moving to new places or give them up as mobile phone and other means of communication have increasingly become popular. As a result, random polls are often exposed to non-response bias mitigating the probabilistic approach of RDD. Nowadays, classical representative polling is less reliable, a trend that will rather continue than cease. Unsurprisingly, lacking representativeness of surveys has been identified as a core reason for recent election polling failures, e.g. in the UK General elections 2015 (Mellon and Prosser 2015).

Can non-representative online polls fix this problem? Since the famously failure of the Literature Digest poll in the 1936 U.S. presidential election, pollsters have been sceptical of non-representative polling (Squire 1988; Goel, Obeng, and Rothschild 2017) and this scepticism is still widespread. Yeager et al. (2011), for example, argue that phone surveys are still more accurate than online polls. However, they base their argument on simple correction approaches for non-probability sampled online polls.[1]

W. Wang et al. (2015) in contrast are much more optimistic about the possibilities of non-representative polling. They used polling results from XBox users, which were highly unrepresentative of the population, to forecast the 2012 U.S. presidential elections. By employing a sophisticated multi-stage approach to post-stratify and calibrate the data they were able to generate accurate forecasts. However, their methodology is mainly suited for large data sets which have sufficient observations for the combined sub-groups (strata) in the sample. Goel, Obeng, and Rothschild (2017) showed that smaller non-representative online surveys can also be accurate. They conducted polls on Amazon Mechanical Turk and a mobile phone app and achieved a level of accuracy sufficient for most practical applications. These works show that online surveys can be used in a meaningful way, if appropriate methods are used to stratify and calibrate the

[1] Goel, Obeng, and Rothschild (2017 Footnote 2) where able to get more accurate results using the basic approach, but in a different way.

non-representative data. Still, how such methods perform for the case of polling and election forecasting remains contested.

3 Data Sources and Quality

3.1 The Dalia Research Europulse poll

In this paper, we are using data from Dalia Research, an online polling firm which is conducting market and opinion research exclusively through smartphones. To ensure to collect data from a broad variety of target populations, Dalia is using a diverse set of app and website categories such as sports, news, entertainment or games. To control that participants answer the survey seriously, an algorithm analyses the consistency and the response behaviour and assigns a "trust score" to every respondent. Dalia praises its methodology as distinctively accounting for potential biases such as interviewer effect, social desirability bias or interviewer data entry errors. (Dalia Research 2016)

Our forecasting project utilizes data of Dalia Research's Europulse Survey which is conducted quarterly in all EU countries. The survey consists of seven waves, but for this project we only use two waves of the survey from December 2016 and March 2017. The first wave is freely available on Kaggle, the second wave was provided to us directly by Dalia Research.[2] Each wave consists of about 11000 individuals, of which roughly 1900 were from Germany which is the fraction of respondents we will focus on in the following analysis. The data is already pre-stratified by Dalia Research based on micro census data for age and gender.

The Europulse data is not particularly collected for election forecasting purposes but contains data on a variety of questions such as online behaviour, media consumption and personal views on political and societal development in the European Union and the respondent's country of origin. Moreover, the survey contains information on the respondent's personal background, demographic data and his or her financial situation. These data can be utilized in order to improve the representativeness of the survey through weighting. This will be explained more detailed in the methodology section.

The election-related variables collected in the Europulse survey are similar to traditional vote intention polling questions. First of all, the respondents are asked if and for which party they will vote in the upcoming election. Moreover, they are asked to rank affiliation with political parties and describe the degree of certainty to cast a ballot for them. Important for our analysis is a variable containing self-reported voting decision at the past German federal election. Within our direct approach, we exploit this data to compute survey weights that adjust the poll to demographic voter

[2] The data for the first wave was collected between December 5th and 15th, 2016. The second wave was conducted between March 13th and 27th, 2017.

composition as reported by the 2013 election statistics.

3.2 Raw forecasts and data quality

As with all surveys, the methodology of Dalia Research has several sources of potential biases (Groves et al. 2009). In case of the Europulse survey they are primarily from flawed measurement and representation of the surveyed population (see figure 1).

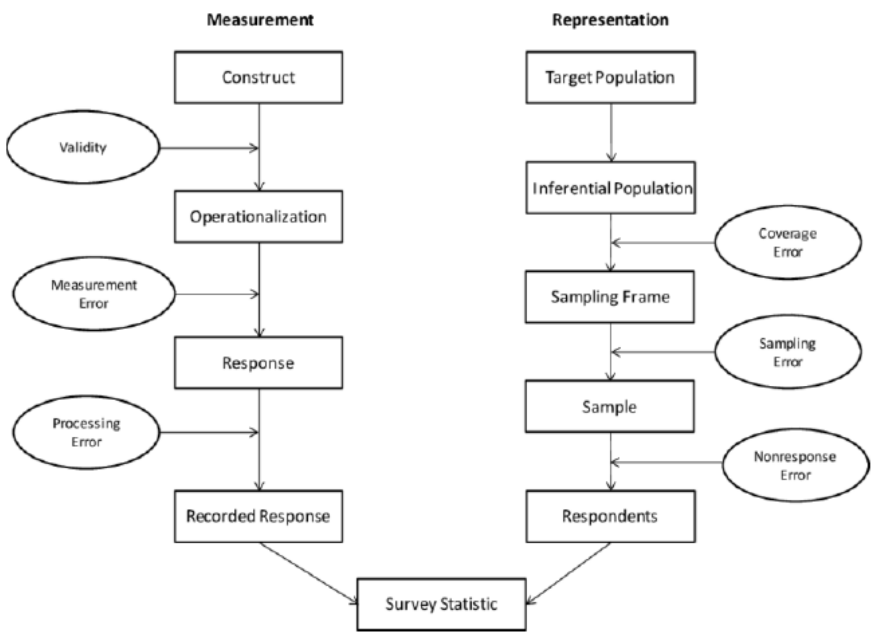
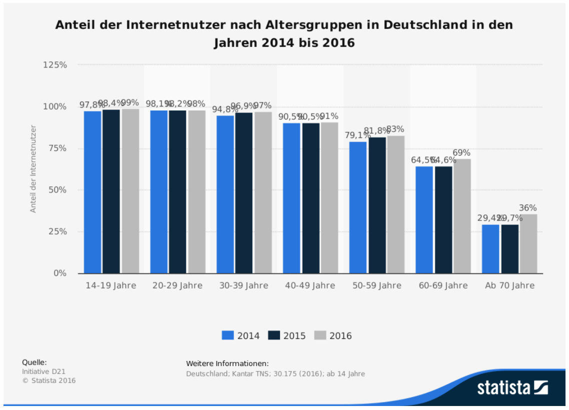


Figure 1: Potential sources of survey error

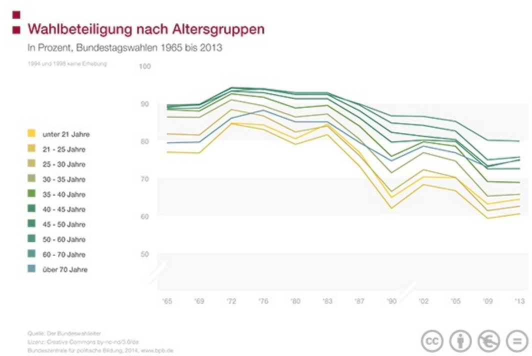
With regard to measuring vote intention for a particular party, the Europulse survey uses a similar approach as traditional surveys: it asks respondents directly which party they intent to vote for at the upcoming election. Whether such questions measure correctly the actual voting behaviour at the election day is questionable but the approach does not differ significantly from traditional polling methods.[3] However, online surveys, such as Europulse, can reasonably claim to avoid some sources of measurement error such as social desirability bias or interviewer bias. Since such surveys are often anonymous, the social pressure on the respondent is presumably negligible, and interaction with the interviewer does not bias the response.

Regarding representativeness, the Europulse survey has some limitations in comparison to face-to-face or RDD interviews. First of all, the Europulse framework does not select participants at random but offers visitors of certain websites or app-users the opportunity to participate. Hence,

[3]The exact question used for instance by *infratest dimap*, a major German polling firm, is: "What party would you vote for if there would be a federal election next Sunday." (dimap 2017) This wording is slightly different than in the Dalia questionaire but evaluating the impact of the formulation of questions shall not be content of this study.



(a) Internet usage by age



(b) Voter turnout by age

Figure 2: Internet usage and turnout by age

this approach carries the risk of self-selection. Moreover, probability sampling would require that Europulse would be generally available to the entire population. However, not everyone is using smartphones and even if they are, they might not use the applications Dalia targets for its surveys. The users Dalia Research actually tends to reach are likely to be much younger and technology oriented than the general population would be (see figure 2). This is in particular a problem for election polling, as older voters are systematically underrepresented in such online methodologies while turnout amongst the elderly tends to be structurally higher.

Dalia Research tries to account for these problems. First, as we have mentioned, they try to increase the diversity of their respondents by presenting their surveys on various apps and platforms targeting different user groups. Second, they pre- and post-stratify their data. Pre-stratification is done on the basis of age and gender, using self-reported demographic information. For each age and gender strata they target a certain number of respondents to adjust their sample to resemble the structure of the German population. In a second step, they use data from the German census and compute weights for cluster of age, gender, education and whether the respondent lives in an urban or rural area. The methodology how Dalia obtained these weights was not available to us. For that reason, we used our own approach for post-stratification that specifically addressed our purpose of election forecasting. Finally, Dalia tries to reduce selection bias since they do not inform respondents that they are answering an election related survey before they actually start it. In general, Dalia Research has high completion rates, hence participants are only rarely dropping out after they started a survey.

Despite these efforts, it is questionable whether the data can be regarded as representative and feasible for polling. In figure 3 we present a simple frequency table of the self-reported past vote in the German federal election 2013 showing that the Dalia sample significantly diverges from the actual voting result.

This large divergence can be due to a variety of flaws that bias Europulse. First of all, even if the

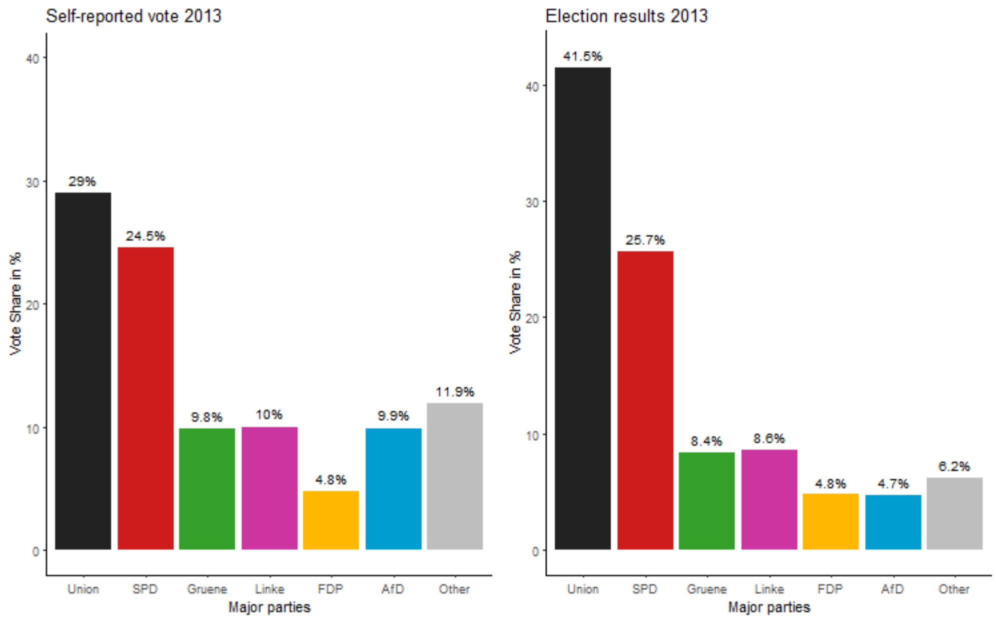


Figure 3: Self-reported vote and actual result 2013

sample represents the German population in terms of age groups and gender, it can be questioned whether the old female or male users are truly representative of their age group. Moreover, not revealing the survey content might sort out politically motivated respondents but interest in political matters is likely correlated to general willingness to respond to surveys. Further problems with the raw data are indicated by the high share of past votes for other parties. This could be a sign of unwillingness to report the true vote intention or unfamilarity with the German party system. Hence, in order to utilize the data to election forecasting further weighting and post-stratification will be necessary.

3.3 Official election statistics and census

The data used for post-stratification and weighting come from several sources. One part of our data is obtained from the German official election statistics (Der Bundeswahlleiter 2017). It contains combined information of turnout and voting decision by gender and age. Age is clustered in groups of 18-25, 26-35, 36-45, 46-60 and 60 plus. This data is the closest estimate of the actual voting population available and is a common measure to account for likely voter bias.

To adjust our sample to demographic composition of the German population we use data of the German Zensus 2011. The data is freely available and we obtained combined frequencies for age, gender and religion. The census claims to reflect the actual demographic distribution of the German population and hence is suitable in order to post-stratify our sample to demographic criteria. Both, election statistics and census is the basis on which we adjust our raw data to obtain representative results.

Finally, in order to set our forecasts into context, we will use polling data from other institutes. Based on the idea of polling the polls (Jackman 2005) we benchmark our adjusted election forecast with the rolling average of other polls. The daily newspaper *Süddeutsche Zeitung* has recently started to publish aggregated polling data similar to Nate Silver's fivethirtyeight.com. We are using their data as they have made it publically available on GitHub.

Finally, as well as the Dalia data, the weighting data suffers from a variety of problems. Most important, the exit poll data we use to compute our weights is not clustered for voters of the *Alternative für Deutschland* (AfD). In order to compute the vote share among groups of age and gender, we took the share of the total election results that the AfD got among all parties that are summarized among *Others*. The weak spot of this approach is that it assumes an equal distribution of AfD voters across all groups of age and gender. While this assumption is very unlikely it is the best approximation we can get with the data at hand.

Moreover, the census as well as the election statistics do not provide combined frequencies of more than three dimensions of demographic variables. This makes it impossible to weight the data on a large set of criteria. Yet, since our Dalia dataset is too small to compute more than three-dimensional cluster anyways that has no severe impact on our methdology.

4 Methodology

We want to employ two approaches to generate election forecasts from our data. In the first approach, the survey data is adjusted in a two step procedure: First, the representativeness of survey data to the general population is increased by finding appropiate weights. Then, a likely voter model is used to obtain an election forecast. Our second approach does this in one step: The survey data will be directly adjusted to the likely voter population based on exit poll data from the German federal election in 2013.

4.1 Post-stratifiation

In order to follow both procedures, we need methods to compute weights for different strata of the survey sample aiming at increased representativeness. A classical way to get these weights is *raking*. With raking weights are assigned to each respondent in order to match the marginal distribution of characteristics in the 'true' population. For example, if we know the distribution of education level and employment status in the population we can compute a weight for each so that the weighted survey sample has the same distribution of the characteristics as the 'true' population. Basically, raking estimates a joint distributions for each combination of characteristics.

However, raking is not ideal. It is often used if only marginal probabilities are available. However, if the joint distribution of characteristics is known we can use *post-stratification* instead. In contrast to raking, post-stratification takes the distribution of combinations of characteristics in the population into account. In other words: It sub-divides the population into groups according to their characteristics. Formally, the post-stratification estimator can be expressed as follows (Goel, Obeng, and Rothschild 2017, Lohr (2010, 142f.)):

$$\hat{y}^{post} = \frac{\sum_{j=1}^{J} N_J \hat{y}_j}{\sum_{j=1}^{J} N_J} \text{ where } \hat{y}_j = \frac{1}{n_J} \sum_{i=1}^{n_J} y_i, \tag{1}$$

where $\hat{y}^{post}$ is the estimator of y in the strata j, N_J the size of the j-th strata in the population and n_J the size of the j-th strata in the sample. From the equation it is obvious that post-stratification is only feasible if we have at least one observation for each strata ($n_J > 0$). But even for small strata sizes in the sample ($n_J < 30$) the standard errors will be large. This is a serious issue as the number of strata grows rapidly with each characteristic included. If we use for instance two gender categories and four age groups we have eight strata, if we add past vote (7 categories) we already have 54. As a result, we might only have a few individuals for each strata or even none. This has severe consequences: For example, estimating the population frequency of the strata female, old (60+) and FDP voter in 2013 with only few respondents in the sample will come with a large error.

Furthermore, our effort to post-stratification comes with another problem: Publically available data often only includes the distribution of pairs of characteristics or, in rare cases, triples (e.g. age, gender and past vote). Yet, the ideal would be to have the frequencies of strata of all variables that is correlated to the voting behaviour of a person. These would be certainly gender and age but also employment status, income, education and religion.[4] Knowing this distribution would enable us to fine tune our forecast. However, as they are not available we will concentrate on combination of characteristics where we have the joint distribtutions. These are in particular age, gender and past votes (from exit polls) as well as combinations of age, gender and religion from the census.

The prior mentioned problem of empty strata can be address in two ways: *Filling strata by aggregating* or with *model-based post-stratification*. Aggregation works simple: If a strata is empty, it has to be combined along one variable with another one. For example: In the December 2016 data set there was no female women above 60 who voted FDP in 2013. In order to address this problem, the age groups 45-59 and 60+ were combined. This is ad-hoc, but a common practice to address empty strata. There is one alternative: In the model-based post-stratification approach the estimates for each strata are not based on the average in the strata as in the equation above, but the result of a multinomial logistic regression. In order to arrive at this regression results, demographic variables in the sample can be used. We did not implement this because the necessary packages for this are not yet released by Goel, Obeng, and Rothschild (2017).[5]

[4]Formally speaking, we need to know the number of individuals in each strata N_j

[5]They are developing an R package for this (postr) and already anounced to make it public.

4.2 Likely voter models

By the time we post-stratified and weighted our sample to resemble the population (in the indirect approach weighting with census data), we have not yet accounted for the problem of likely voter bias. Surveys tend to either include voters who say they will vote but eventually do not do so, or include voters who say they will not vote but finally cast a ballot (Bernstein, Anita, and Montjoy 2001). Since this behaviour of individuals has turned out to be correlated across party preferences it is likely to bias our forecast (Keeter, Igielnik, and Weisel 2016).

Two families of methods are proposed by Keeter, Igielnik, and Weisel (2016) to account for likely voter bias - *deterministic* and *propabilistic* methods. The former was initially proposed in Perry (1960) and Perry (1979) and tries to identify criteria that describe a voter as a likely voter on a scale from 0-7. Using a distinct set of questions such as "How likely are you to vote at the next national election?" respondents are assigned points on the likely voter scale and are eventually ranked. Using an estimate of the upcoming election turnout, a cut-of criteria determines which rank on the likely voter scale is needed to be included in the sample measuring the actual voting population.

The downside of such methods is that people below the threshold value of the scale have no affect at all on the actual forecast. This is addressed by probabilistic methods. Such an approach estimates on the basis of a set of predictor variables like demographic characteristics, partisanship or ideology a probability of a person to cast a ballot. By using this probability even people with a very low likelihood of voting affect the actual estimate to a certain degree. The downside of such methods is that it requires data on turnout and the predictor variables of the past election. Moreover, such models assume that turnout is time-stable for the population groups across past and future elections.(Keeter, Igielnik, and Weisel 2016)

For the purposes of our calibration we decided to use the simplest deterministic approach: Self-reported vote intention. For further fine-tuning we could employ other approaches, but our analysis (see below) showed that for our data set the post-stratification is more crucial.

5 Results

Our results show that the choice of weighting method has a considerable effect on the performance of the forecast against our benchmark data. Moreover, we could show that proper and transparent weighting with publicly available elections statistics and census is feasible and leads to representative and reasonable forecasts from a non-representative online poll. Table 1 displays how the raw and the adjusted results for December 2016 and March 2017 perform in comparison to the *Süddeutsche Zeitung's (SZ)* rolling average using the root mean squared error (RMSE) as indicator of accuracy.

As the table shows, the divergence of the raw data was relatively high compared to the average

Table 1: Raw, weighted and benchmark estimates

December								
Method	Union	SPD	FDP	Gruene	Linke	AfD	Other	RMSE
SZ Rolling Average	35.50	22.20	5.70	10.60	9.70	11.60	4.70	
GAV Exit Polls	32.50	21.10	8.70	8.60	9.60	11.70	7.80	2.17
GAR Census Data	24.60	21.60	5.50	10.60	9.50	14.00	14.20	5.55
Dalia Unweighted	23.80	20.90	6.10	10.50	10.50	13.80	14.40	5.83
March								
Method	Union	SPD	FDP	Gruene	Linke	AfD	Other	RMSE
SZ Rolling Average	33.10	31.40	5.80	7.60	7.70	9.30	5.10	
GAV Exit Polls	36.70	28.10	4.90	7.50	8.60	7.80	6.40	2.05
GAR Census Data	23.10	30.00	5.80	8.40	9.00	13.80	10.00	4.61
Dalia Unweighted	23.60	28.40	5.60	7.70	9.80	14.50	10.30	4.75

Notes: The benchmark is the rolling average of all German major polls as computed by *Süddeutsche Zeitung*. GAV stands for strata combined of the variables gender, age and self-reported vote at the last election. GAR indicates strata of gender, age and religion. The last row of each month represents the raw results without weighting as collected by Dalia Research.

of official polls. The raw forecast has a RMSE of 5.83 for December and 4.75 in March. Yet, with our weighting approach using census data we were able to improve this performance only by a negligible amount. Employing clusters of gender, age and religion our forecasts show a RMSE of 5.55 and 4.61 respectively. The weak performance might be due to several reasons. As stated earlier, the Dalia data is pre-stratified with age and gender by the time the poll is conducted. We decided, however, to account for age and gender through census data since we did not transparently evaluate Dalia's procedure for stratification. Yet, the very low difference of our estimate with Census data and the original data shows that our approach did not siginificantly differ from Dalia's approach. Moreover, the variable religion does not seem to have a reasonable impact on our weights, which is also partially a result of the fact the religious groups and the Dalia religious category do not perfectly overlap. The census counts religious communities according to German public law (Islam is not one) while the Dalia survey askes for self-reported religious affiliation. We adjusted the post-stratified data for the likely voter, but this had only a relatively small impact on the vote shares.

In contrast, our approach using the election statistics does perform very well compared to census data. Weighting with cluster of age, gender and vote decision and taking into account self-reported vote at the past election yields results in RMSE values of 2.17 and 2.05 respectively compared to our benchmark. Figure 4 contrasts the raw data with our weighted estimates for December and March.

These results are very promising. They show the possibility of adjusting online polling data with exclusively public available information in a way that they converge to professionally conducted traditional polls. While our results are not conclusive to assess the usefulness of online polls for election forecasting in general, we can, however, state that our two weighted samples are doing well. In addition, looking on the time dimension yields further insights. Figure 5 shows the trend

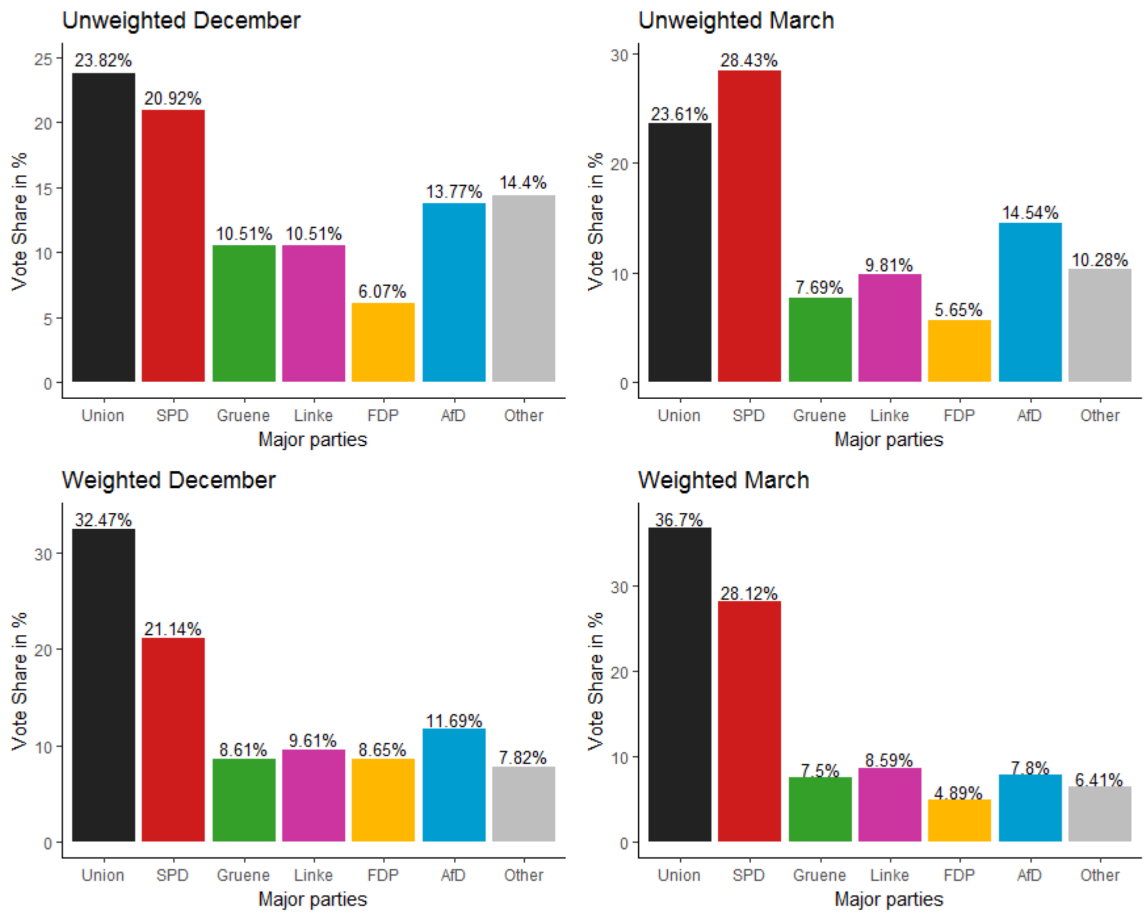


Figure 4: Weighted and unweighted polls

of the poll's rolling average since 2016. The dots display our adjusted data.

This shows, how close our data is to the benchmark and that it picked up the large shift in vote shares for the SPD - the so called "Schulz-Effect" - after the annoucement of Martin Schulz as chancelor candidate. Moreover, our estimates are largely within the confidence intervals of the aggregated polls (colored areas below and above the rolling average line). While this illustrates that the online polls adjusted by age, gender and last vote are a significant improvement to the raw data there is still considerable room for improvement. This could be achieved either by using more detailed clusters, different approaches to account for likely voter bias or other data sources such as exit polls of different polling institutes.

6 Conclusion

In this paper, we tested different adjustment methods for online election polls conducted through smartphone apps. The data is highly unrepresentative of the voting population and yields unrea-

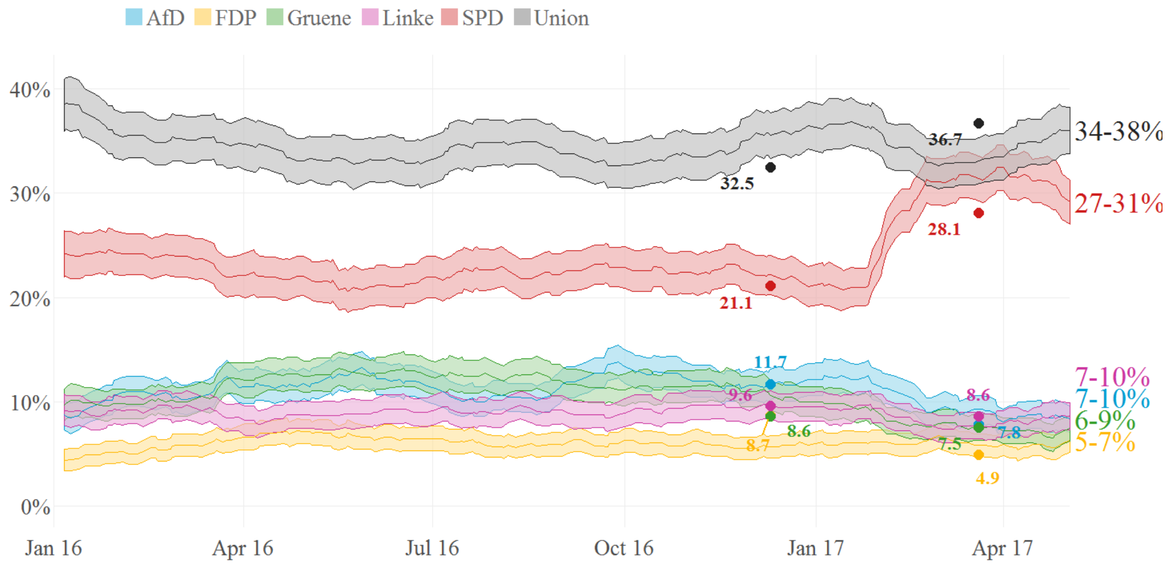


Figure 5: Weighted polls benchmarked to overall poll's rolling average

sonable results in its raw format. By using different post-stratification methods we calibrated the polls and produced reasonable election forecasts. We find that weighting with self-reported past vote is a promising approach to correct online polls, while our purely demographic approach did not perform well. But this result might be the consequence of choosing the wrong demographic variables: Age and gender were already considered in pre-stratification and religion had a measurement problem. A possible alternative would be to use other variables for post-stratification, e.g. education, employment status or information on the urbanization which are considered by Dalia for pre-stratification. Our second approach, weighting with clusters of age, gender and self-reported voting decision in 2013 against exit poll data, yielded weights that performed relatively well compared to aggregated polls. We conclude that exit poll data is useful to account for voting population representativeness and likely voter bias.

The shortcomings of our approach were twofold. First, since the exit polls do not report AfD voters across clusters of age and gender, we had to impute this data from the aggregated data of parties summarizes among *Others*. This induced error and probably reduced the accuracy of our AfD forecast. Second, our approach assumes that the clusters of actual voters remain constant from 2013 to 2017. But introducting more complex assumptions on the characteristics of likely voters requires more sophisticated methods of identification. Such 'likely voters models' as explained in chapter 4.2 and the employment of more detailed cluster of exit poll data would presumably make our polls even more accurate. However, our primary aim was to show that non-representative online polls can be adjusted to produce reasonable results. In order to publish the forecasts, further fine-tuning would be required.

References

Bernstein, Robert, Chadha Anita, and Robert Montjoy. 2001. "Overreporting Voting: Why It Happens and Why It Matters." *Public Opinion Quaterly* 65: 22–44.

Dalia Research. 2016. "Dalia Research Methodology." https://daliaresearch.com/wp-content/uploads/2016/08/Methodology-PDF-1.pdf.

Der Bundeswahlleiter. 2017. "Ergebnisse Der Representativen Wahlstatistik." https://www.bundeswahlleiter.de/bundestagswahlen/2013/ergebnisse/repraesentative-wahlstatistik.html.

dimap. 2017. "Allgemeine Informationen Zur Sonntagsfrage." infratest dimap. https://www.infratest-dimap.de/service/faqs/faqs-zur-sonntagsfrage/.

Gallup. 2010. "Understanding Gallup's Likely Voter Models." http://www.gallup.com/poll/143372/understanding-gallup-likely-voter-models.aspx?version=print.

Goel, Sharad, Adam Obeng, and David Rothschild. 2017. "Online, Opt-in Surveys: Fast and Cheap, but Art They Accurate?" Working Paper.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2nd ed. New Jersey: John Wiley & Sons Inc.

Holbrook, Allyson L., Jon A. Krosnick, and Alison Pfent. 2007. "The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms." In *Advances in Telephone Survey Methodology*, edited by James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, 499–528. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9780470173404.ch23.

Jackman, Simon. 2005. "Pooling the Polls over an Election Campaign." *Australian Journal of Political Science* 40 (4): 499–517. doi:10.1080/10361140500302472.

Keeter, Scott, Ruth Igielnik, and Rachel Weisel. 2016. "Can Likely Voter Models Be Improved? Evidence from the 2014 U.S. House Elections." Pew Research Center.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National Rdd Telephone Survey." *The Public Opinion Quarterly* 70 (5): 759–79. http://www.jstor.org/stable/4124225.

Lohr, Sharon. 2010. *Sampling: Design and Analysis*. Brooks/Cole.

Mellon, Jonathan, and Chris Prosser. 2015. "Investigating the Great British Polling Miss: Evidence from the British Election Study." *SSRN Electronic Journal*. doi:10.2139/ssrn.2631165.

Perry, Paul. 1960. "Election Survey Procedures of the Gallup Poll." *Public Opinion Quaterly* 24:

531–42.

———. 1979. "Certain Problems in Election Survey Methodology." *Public Opinion Quaterly* 43: 312–25.

Pew Research Center. 2012. "Assessing the Representativeness of Public Opinion Surveys." Pew Research Center. http://www.people-press.org/files/legacy-pdf/Assessing%20the%20Representativeness%20of%20Public%20Opinion%20Surveys.pdf.

Skibba, Ramin. 2016. "The Polling Crisis: How to Tell What People Really Think." *Nature* 538 (7625): 304–6. doi:10.1038/538304a.

Squire, Peverill. 1988. "Why the 1936 Literary Digest Poll Failed." *The Public Opinion Quarterly* 52 (1): 125–33. http://www.jstor.org/stable/2749114.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91. doi:10.1016/j.ijforecast.2014.06.001.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang. 2011. "Comparing the Accuracy of Rdd Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75 (4): 709–47. doi:10.1093/poq/nfr020.