Election Forecasting using Postratification Methodology on Dalia Research's Europulse Data*

Moritz Hemmerlein & Alexander Sacharow

18 July, 2017

Summary or TL;DR

- For election forecasting Dalia's Europulse data faces similar problems as conventional polling data; i.e identification of likely voter (representativeness for actual voting population)
- Conducting post-stratfication with official election statistics ("exit polls") and information on self-reported voting behaviour at the last election, we obtained fairly accurate results
- The weighted Europulse poll yielded results close to the moving average of leading German polling institutes (RMSE of 2.05 for the March data)
- There is room for further improvement of the weighting algorithm as well a usage of other benchmarks to evaluate the accuracy of the estimates

*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.

Introduction

Using online polling data for election forecasting is fairly unexplored and required some initial reflection on advantages and shortcommings compared to traditional approaches. Analyzing potential sources of error shows that both approaches are not fundamentally different. While traditional polls are pre-stratified and post-stratified to ensure representativesness, online polling data is post-stratified using similar demographic information.

We decided to employ conventional weighting methods used traditonally to make election forecasts on the Europulse online data. Our hypothesis was that the online data weighted in a similar fashion should not perform consiberably worse than conventional data when it comes to polling. Our analysis supported this claim. Benchmarking our estimates with the moving average of traditional polls the margin of error turned out to be fairly low.

While we conducted some different approaches it turned out that one approach performed much better than the others. We will present solely this most promising approach in this summary.

Methodology

weighting method

variable selection -> we chose xy, others can be chosen

Data

The data used for post-stratification and weighting comes 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.

The big shortcomming of the data is that it is available in no more than two-dimensions, which means that the data is never clustered by more than two demographic variables. Potentially, other and more detailled exit poll data could be used to improve the post-stratification and further enhance accuracy and representativeness for actual voter.

As benchmark data we used the rolling average as computed by the *Süddeutsche Zeitung*. The raw data and the computing method can be found on GitHub.

Results

Following, we display our results for the December and the March data respectively. The first figure shows the election forecast yielded by the unweighted raw data and benchmarks it with the weighted results. As the graph shows, the forecasts differ significantly. The raw data has a strong bias towards the smaller parties and non-voters. Weighting this data yields more reasonable results with improvements for the CDU/CSU and the SPD while the smaller parties loose. Moreover, the expectation of non-voters gets more accurate.

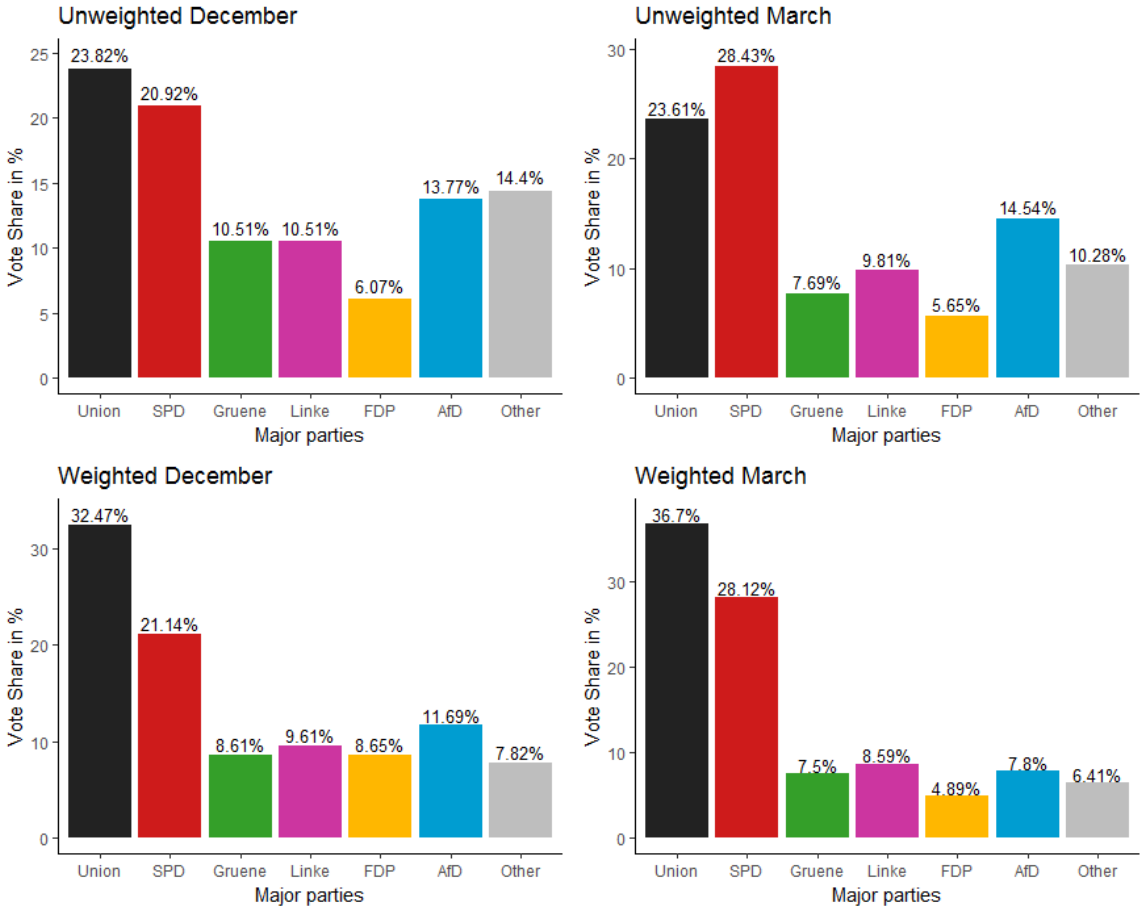

Figure 1: Weighted and unweighted polls

The table below summarises the numerical results and displays the the root-mean-squared error of the estimates with the benchmark data (polling institutes rolling average). As mentioned before, the weighting by gender and age using exit poll data yields a RMSE of 2.17 and 2.05 for December and March respectively.

Graphically the results are shown in the graph below. As one can see, the weighted estimates (points in the graph) move smoothly with the moving average forecast that is used as a benchmark.

Moritz Hemmerlein
Alexander Sacharow

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.

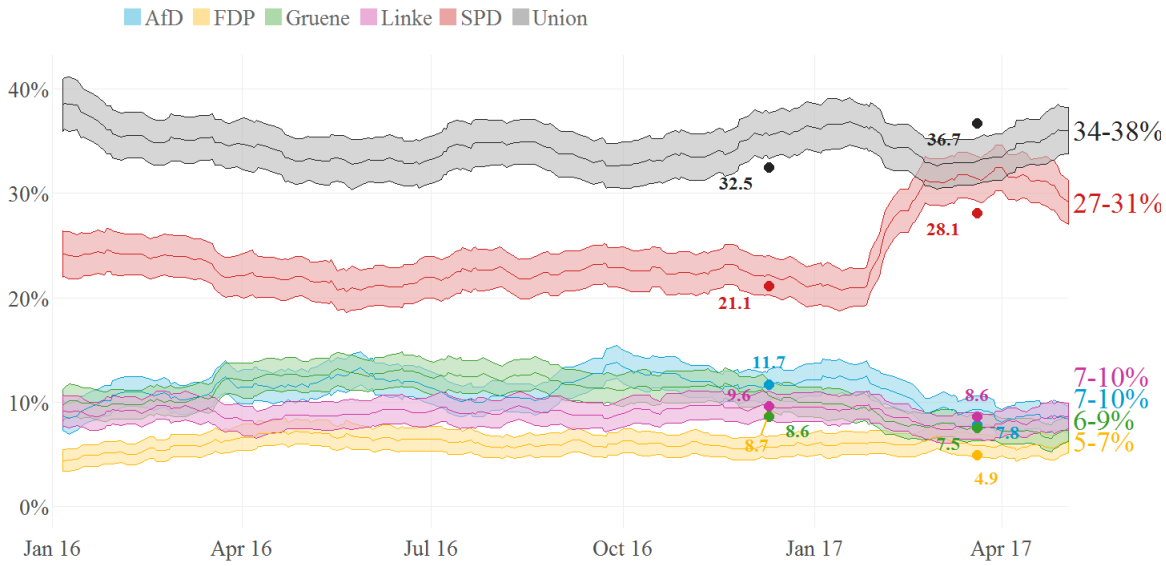

Figure 2: Weighted polls benchmarked to overall poll's rolling average

Shortcomings and Potential for Improvement

- more detailed stratification
- other demographic information
- improving the questionaire to identify likely voters

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