Forthcoming (subject to changes) in the

International Journal of Forecasting

Accuracy of German federal election forecasts, 2013 & 2017

Andreas Graefe
Business School, Macromedia University, Munich, Germany
Graefe.andreas@gmail.com

Abstract

The present study reviews the accuracy of four methods (polls, prediction markets, expert judgment, and quantitative models) for forecasting the two German federal elections in 2013 and 2017. On average across both elections, polls and prediction markets were most accurate, while experts and quantitative models were least accurate. The accuracy of individual forecasts did not correlate across elections. That is, methods that were most accurate in 2013 did not perform particularly well in 2017. A combined forecast, calculated by averaging forecasts within and across methods, was more accurate than two out of three component forecasts. The results conform to prior research on US presidential elections in showing that combining is effective in generating accurate forecasts and avoiding large errors.

Keywords

Combining forecasts, election forecasting, polls, prediction markets, expert judgment, econometric models

1 Introduction

Combining forecasts has a long history in forecasting research (Bates & Granger, 1969). Numerous studies have shown that combining reduces forecasting error (Clemen, 1989), particularly when one draws upon forecasts from different methods that use different data (Armstrong, 2001).

Combining forecasts is beneficial for a variety of reasons. First, any single forecasting method or model can only include a limited amount of information (or data). As a result, the resulting forecasts may lack information and could thus be subject to bias. Combining forecasts from different methods and models allows for incorporating more information and can thus help to reduce bias. Second, when using simple unweighted averages to combine forecasts, the combined forecast will always be at least as accurate as a randomly chosen component forecast. What is more, a combined forecast often outperforms even its most accurate component. Third, forecasts that rely on different methods and / or data are more likely uncorrelated and therefore are more likely to "bracket" the final outcome. If bracketing occurs, the combined forecast cancels out some of the systematic and random errors of individual forecasts, and thus reduces error. Third, the relative forecast accuracy of different methods often varies across time, with no relationship between past and future accuracy. In other words, methods that provided highly accurate forecasts in the past might be among the least accurate methods when predicting the future. When relying on a combined forecast, forecasters avoid the danger of picking a poor forecast. (Graefe, Armstrong, Jones, & Cuzán, 2014b).

The present study provides new empirical evidence from comparing and combining forecasts for German federal elections. This work is part of the PollyVote project, which was founded in 2004 with the goal to apply general findings derived from forecasting research to election forecasting. The PollyVote is a long-term project. To this day, the PollyVote method has been used to forecast the four U.S. presidential elections from 2004 to 2016 (Graefe, Armstrong, Jones, & Cuzán, 2017), as well as the German federal elections in 2013 (Graefe, 2015) and 2017 (i.e., the present study). In applying the method to many elections over time and in different countries, the aim is to learn more about relative accuracy of different election forecasting methods, and combinations thereof.

2 Combining forecasts

Combining forecasts is valuable whenever one has more than one forecast available to predict a particular outcome. Yet, ideally, forecasters have access to (1) many validated forecasts that (2) draw upon different methods and data (Armstrong, 2001). And, (3) they are unsure as to which forecast will provide the most accurate predictions.

The case of election forecasting meets these conditions (Graefe et al., 2014b). First, there are many evidence-based methods for predicting election outcomes, such as polls (Pasek, 2015), prediction markets (Graefe, 2017a), expert judgment (Jones & Cuzán, 2013), citizen forecasts (Graefe, 2014), and quantitative models (Lewis-Beck, 2005). Second, these methods rely on different data. Polls feed of people's vote intentions. Prediction markets, expert judgment, and citizen forecasts harness people's expectations of who is going to win, whereas quantitative models predict election outcomes from structural "fundamental" data (e.g., economic indicators or the time the incumbent party has been in power). Third, as for most forecasting problems, it is very difficult to determine a priori which method will eventually turn out to be the most accurate.

While the first two conditions are easy to understand, many people have difficulties accepting the third condition. This is because people commonly think that they know which forecast (method) is the best and decide to follow it. For several reasons, this is not a good approach to forecasting. First, 'cherry-picking' forecasts often leads to selecting a forecast that suits one's own biases, and ends up being less accurate that the combined forecast (Soll & Larrick, 2009). Second, in most practical situations it is

extremely difficult – if not impossible – to know in advance which forecast will prove to be the most accurate. Past accuracy, for instance, is not a good predictor of future accuracy: two studies in the domain of election forecasting found a negative correlation between the predictive and historical accuracy of a forecasting method (Graefe et al., 2017) or model (Graefe, Küchenhoff, Stierle, & Riedl, 2015). Third, assume for a moment that it would be possible to identify the most accurate forecast in advance. Even then, combining that forecast with another (less accurate) forecast can be a good strategy, as illustrated by Herzog and Hertwig (2009). The authors showed that an equally weighted average of two forecasts outperforms the best individual forecast if (a) the two forecasts enclose the outcome and (b) the error of the less accurate forecast is not more than three times the error of the more accurate one. From a practical viewpoint this means that the error of an additional forecast that is added to the combination can be quite large, as long as it increases the chance that the range of forecasts captures the true value.

The question of *how* one should combine forecasts is uncritical. Despite researchers' vast efforts to develop and test sophisticated weighting methods for combining forecasts, the simple average remains difficult to beat, and often provides more accurate forecasts than complex approaches to combining (Genre, Kenny, Meyler, & Timmermann, 2013; Graefe et al., 2015). One reason for the accuracy of the simple average is, as mentioned above, that the relative accuracy of component forecasts varies over time and is difficult to predict. In such a situation, weighting the forecasts, for example based on past performance, is of limited value.

3 The combined PollyVote forecast

The PollyVote follows a two-step approach for combining forecasts within and across different forecasting methods, the so-called component methods, each of which rely on different data. First, forecasts are averaged *within* each component method before, second, averaging the resulting forecasts *across* the component methods (cf. Figure 2). This section briefly reviews the PollyVote's predictive accuracy in previous elections.

3.1 U.S. presidential elections

The PollyVote has been used to forecast the popular vote in the four U.S. presidential elections in 2004 (Cuzán, Armstrong, & Jones, 2005), 2008 (Graefe, Armstrong, Jones, & Cuzán, 2009), 2012 (Graefe, Armstrong, Jones, & Cuzán, 2014a), and 2016 (Campbell et al., 2017). In addition, the method has been tested retrospectively for the three elections from 1992 to 2000 (Graefe et al., 2014b).

One study summarized the PollyVote's performance across these seven elections. With a MAE of 1.1 percentage points across the last 100 days before each election, the PollyVote's forecast error was lower than the corresponding error of any other method component method. Error reductions ranged from 8%, compared to citizen forecasts, to 58% compared to polling averages (Graefe et al., 2017).

3.2 German federal elections

The PollyVote has also been used to predict the vote shares of six parties in the 2013 German Federal election by combining forecasts from polls, prediction markets, econometric models, and expert judgment. On average, across the two months prior to the election, which is the maximum time frame for which data were available, the PollyVote provided more accurate predictions than the typical component forecast, with error reductions ranging from 5%, compared to polls, to 41%, compared to prediction markets (Graefe, 2015).

4 Forecasting the 2017 German federal election

The 2017 German federal election was held on Sunday, September 24, to elect the members of the national parliament, the so-called Bundestag. Chancellor Angela Merkel's CDU/CSU won 32.9% of the vote. While this result made the CDU/CSU again the strongest party, the party lost 8.6 percentage points compared to the previous election in 2013. While in 2013, the CDU/CSU achieved their best results since 1990, the outcome of the 2017 election was the party's worst result since 1949. Things were even worse for the Social Democrats (SPD), the junior member of the governing Grand Coalition. The SPD won only 20.5% of the vote, which was the party's lowest vote share since the Second World War. In contrast, the Alternative for Germany (AfD), a right-wing populist party founded shortly before the 2013 election, entered parliament for the first time in their still young history, and became the third largest party with 12.6% of the vote. In addition, three other parties made the 5% threshold to enter parliament, namely the Free Democratic Party (FDP) with 10.7%, the Left with 9.2%, and the Greens with 8.9%.

4.1 Method

The present study analyzes the relative accuracy of different methods that have been used to forecast the 2017 German election.[1] A forecasting method's absolute error on a particular day was calculated by averaging the absolute differences between the predicted and actual vote shares of the six largest parties and the remaining share for all other parties combined (i.e. CDU/CSU, SPD, AfD, FDP, Left, Greens, and Others). When a method did not publish a new forecast on a particular day, the forecast from the previous days was used. Then, each method's mean absolute error (MAE) was calculated across the last 63 days (or nine weeks) prior to the election, as this is the time period when forecasts from all individual component forecasts were available.

4.2 Results

This section describes the accuracy of each component method, and the combined PollyVote forecast, in predicting the 2017 German federal election. For more details and further literature on the various forecasting approaches used in different countries see Stegmaier and Norpoth (2017). All data are will be made available upon publication at the Harvard Dataverse.

4.2.1 Polls

Polls ask people for whom they intend to vote if the election was held today. Thus, polls measure public opinion at a certain point in time; they do not provide predictions of what will happen on election day. Nonetheless, polling results are commonly projected to election day and interpreted as forecasts (Hillygus, 2011). Figure 1.A shows the MAE for polls published by seven established pollsters across the last 63 days before the election. The polling data were obtained from Wahlrecht.de, a website that collects and published polls for German federal and state elections. The MAE ranged from 2.1 percentage points (Infratest dimap) to 2.8 percentage points (Allensbach).

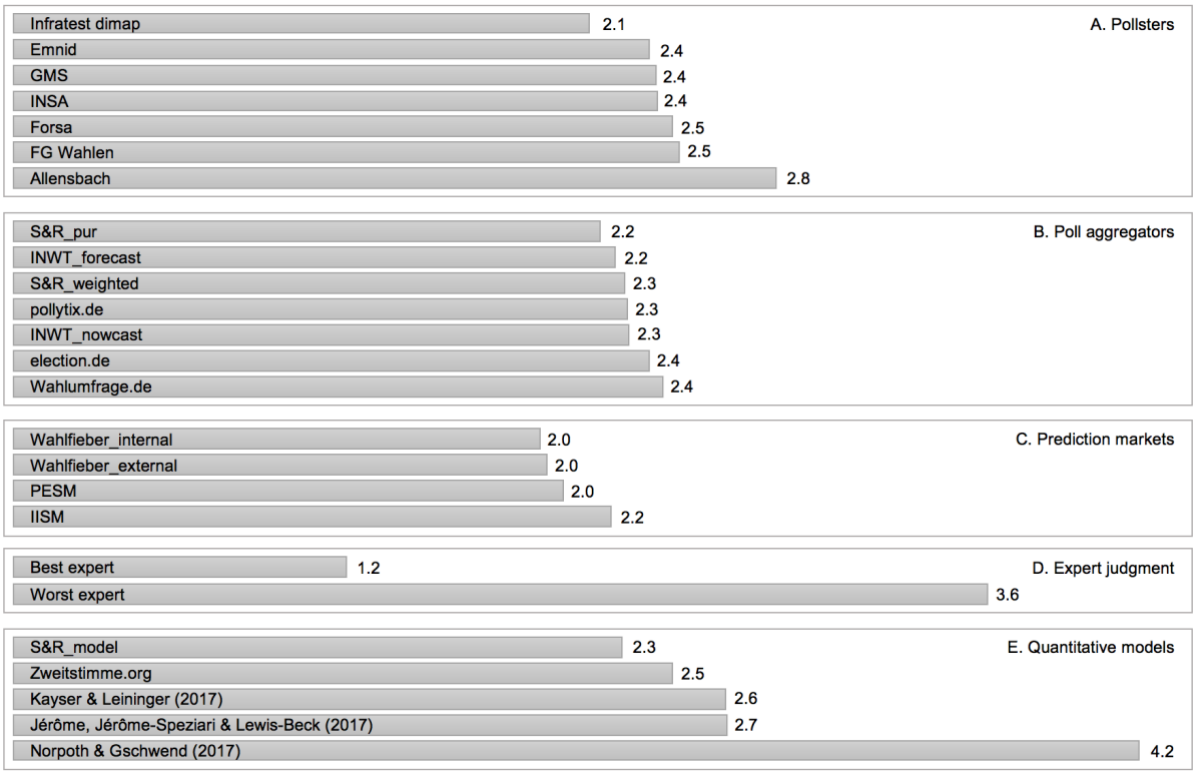
[1] The combined PollyVote and its component forecasts were published at pollyvote.de, and updated on a daily basis through an automated process. The analysis in the present paper, however, is not based on the 'real-time' data. In order to limit the risk of introducing errors from real-time analyses, the present paper uses the final datasets obtained after the election.

4.2.2 Poll aggregators

Online poll aggregators have become increasingly popular in Germany, and several websites reported polling averages prior to the 2017 election. The PollyVote collected data from five websites who published a total of seven different polling averages. As shown in Table 1, the methods for aggregating polls varied regarding which polls were included and how they were aggregated.

- Election.de used the most basic approach by calculating simple unweighted averages of the most recent polls from three survey institutes, namely Forsa, Forschungsgruppe Wahlen (FGW), and Infratest dimap. The polling average was updated once a week.
- Wahlumfrage.de also calculated simple unweighted averages but used polling data from the seven pollsters listed in Figure 1 (Section A). The Wahlumfrage.de polling average was updated only periodically.
- Pollytix.de calculated rolling weighted averages and also included polls from other survey institutes (e.g., YouGov). The PollyTix weighting approach assigned higher weights to surveys with larger samples and to surveys that were conducted more recently.

Figure 1: Mean absolute error for each individual component forecast, calculated across the last 63 days prior to the 2017 election



- Signal & Rauschen (signalundrauschen.de) published two polling averages, hereafter referred to S&R_pur and S&R_weighted.
    a. S&R_pur used the same approach and data as Wahlumfrage.de but updated the polling average every week.
    b. S&R_weighted was more complex in calculating rolling weighted averages of all polls that were published across the previous 20 days. Thereby, more recent polls as well as polls with larger samples received more weight in the combination. In addition, the method accounted for individual pollsters' house biases (i.e., whether a pollster systematically over- or underestimated certain parties in previous elections).

- The consulting firm INWT Statistics calculated rolling weighted averages based on the Wahlrecht.de data by accounting for individual pollsters' house biases as well as their overall accuracy in previous elections. INWT published two versions of their polling average: a nowcast and a forecast. The difference is that the INWT_forecast was more conservative than the INWT_nowcast as it reacted less strongly to recent changes in polling numbers.

As shown in Figure 1.B, differences in accuracy between the various poll aggregators were small. Across the last 63 days, the MAE ranged from 2.2 percentage points for S&R_pur and INWT_forecast to 2.4 percentage points for Wahlumfrage.de. The simple S&R_pur, which calculated unweighted averages of polls from seven established survey institutes, was slightly more accurate than the more complex methods that relied on weighted averages. The four measures that relied on rolling weighted averages performed more or less similar. The two methods that incurred the largest errors either relied on only few pollsters (i.e., election.de) or were updated infrequently (i.e., Wahlumfrage.de). But, again, differences in accuracy were small.

Table 1: Mean absolute error of different poll aggregators (calculated across the last 63 days prior to the 2017 election)

					Weighting methodology				
Poll aggregator	**MAE**	**N of pollsters**	**Frequency**	**Type of average**	**Recency**	**Sample**	**House bias**	**Past accuracy**	**Conservatism**
S&R_pur	2.2	7	Weekly	Unweighted	-	-	-	-	-
INWT_forecast	2.2	7	Rolling	Weighted	No	No	Yes	Yes	Yes
S&R_weighted	2.3	7	Rolling	Weighted	Yes	Yes	Yes	No	No
pollytix.de	2.3	> 7	Rolling	Weighted	Yes	Yes	No	No	No
INWT_nowcast	2.3	7	Rolling	Weighted	No	No	Yes	Yes	No
election.de	2.4	3	Weekly	Unweighted	-	-	-	-	-
Wahlumfrage.de	2.4	7	Periodically	Unweighted	-	-	-	-	-

Weighting methodology
Recency: more recent polls weighted more heavily
Sample: polls with larger sample weighted more heavily
House bias: accounts for house biases of individual pollsters
Past accuracy: pollsters that were more accurate in the past weighted more heavily
Conservatism: damping of dramatic changes in polling numbers

4.3.3 Prediction Markets

Betting on election outcomes has a long history and can be traced back to 16th-century Italy, where such markets were common for civic and papal elections (Rhode & Strumpf, 2014). Long before the emergence of scientific polling, such markets were also popular in the U.S., where leading newspapers such as the New York Times would betting odds as forecasts of what might happen on Election Day (Rhode & Strumpf, 2004).

Prediction markets use the price mechanism of the market to aggregate people's expectations of how the election will turn out. For German elections, prediction markets typically offer contracts for each party, where the contract price reflects that party's predicted vote (e.g., a price of 30 Euros for CDU/CSU means that the party is predicted to win 30% of the vote). Participants who think that the actual vote share will be higher (lower), should buy (sell) shares of that contract, and win or lose money depending on the accuracy of their predictions.

A review of prediction market accuracy of vote share forecasts in different countries found that prediction markets tend to outperform forecasts made by polls, models, and experts; compared to simply asking citizens who will win, evidence was mixed (Graefe, 2017a). However, that review did not include the 2016 U.S. presidential election, a case when prediction markets provided relatively poor forecasts (Graefe, 2017b).

For the 2017 German election, the PollyVote collected data from three prediction market providers, namely PESM, FAZ Orakel (operated by IISM), and Wahlfieber.de, which offered forecasts

from four markets (Wahlfieber operated two markets, an internal one, Wahlfieber_internal, and an open one to anyone to participate, Wahlfieber_external).

The FAZ Orakel and the two Wahlfieber markets operated with play money. That is, participants received a certain amount of play money that they could use for trading vote shares of the six largest parties plus the combined vote of all remaining parties. Traders' performance on play-money markets is measured through rankings and, in some markets, the best performing participants can win prizes. As described by Graefe (2017a), the risk of potential market manipulation is higher for play money markets than for markets operating with real money. At the PESM market, participants traded real money but investment was limited to 10 Euros per participants.

Figure 1.C shows the MAEs for the different prediction markets across the last 63 days before the election. The markets performed very similar. The MAEs of both Wahlfieber markets and the PESM were 2.0 percentage points. The MAE of the remaining market, the FAZ Orakel, was 2.2 percentage points, which is still on par with the most accurate poll aggregator.

4.4.4 Expert Judgment

Expert surveys have a long history as a method to forecast election outcomes (Kernell, 2000). Experts are assumed to provide accurate forecasts due to their domain knowledge. Experts may, for example, be able to correctly interpret polls and project their results to Election Day, by taking into account potential impacts of recent and future campaign events. Some evidence suggests that this is the case. Jones and Cuzán (2013) found that the combined forecast of experts outperformed polls for long-term forecasts. For the four U.S. presidential elections from 2004 to 2016, Graefe (2018) found that the majority of 452 individual experts forecasts in correctly predict the directional error of polls. That said, the typical expert's error was 7% higher than the corresponding error of a polling average. The findings from that study also suggested that experts do not sufficiently harness information incorporated in the so-called fundamentals (see Section 4.4.5).

Prior to the 2013 election, 70 members of the German Society for Electoral Studies (DGfW) were asked to participate in a monthly online survey. The first survey round started in March 2017, and the last survey round was conducted the week before the election. Across all seven survey rounds, a total of 37 experts participated. The average number of experts per survey was 27 and ranged from 25 to 30.

Figure 1.D shows the MAE of the most accurate and least accurate of those 17 experts who participated in all four survey rounds from June 30 to Election Day. The most accurate expert had a MAE of only 1.2 percentage points, which made that person's forecasts substantially more accurate than all other component forecasts. At the other end of the spectrum, the MAE of the least accurate expert was 3.6 percentage points, the second largest error of all other component forecasts.

4.4.5 Quantitative Models

Developing statistical models is an alternative to relying on people's vote intention (i.e., polls) or expectations (i.e., prediction markets, expert judgment) when generating election forecasts. Such models typically use structural factors, the so-called fundamentals, that are predictive of election outcomes. For example, the incumbent government typically benefits from good economic conditions but loses support the longer it has been in power due to people's desire for change (Lewis-Beck, 2005). Forecasts from five quantitative models were available prior to the 2013 election.

- The model by Jérôme, Jérôme-Speziari, and Lewis-Beck (2017) has been used in modified form since 1998. The model provides a one-off forecast of the vote shares of all parties that are represented in the outgoing parliament based on the unemployment rate and several poll-based measures (e.g., the popularity of the Chancellor candidates of CDU/CSU and SPD, the popularity of the FDP as a coalition partner and vote intention for the smaller parties).

- The Chancellor model by Norpoth and Gschwend (2017), which has been around since the 2002 election in modified form, uses three variables to forecast the aggregate vote share of different coalitions: (1) the coalition's average vote share across the three preceding elections, (2) the support for the two major parties' Chancellor candidates in public opinion polls, and (3) attrition, measured as the number of terms in office.
- Zweitstimme.org published a hybrid model that combined a structural component with a poll-based component. The model provided daily updated forecasts of the vote shares of the six largest parties (Munzert, Stötzer, Gschwend, Neunhoeffer, & Sternberg, 2017). The structural component is based on three variables, namely (1) a party's vote share in the previous election, (2) the average party support in polls published 230 to 200 days before the election, and (3) a binary variable to indicate the party of the Chancellor. The poll-based component was an adaption of the random walk model by Linzer (2013).
- The model by Kayser and Leininger (2017), another newcomer in this election, predicts the parties' vote share in each state before aggregating the numbers to a forecast of the federal election outcome. The linear random effects model is based on the following information: the party's vote share in the preceding federal and state election, whether the Chancellor was from that party, national quarterly GDP growth, and the number of years the chancellor has been in office. The authors provided two sets of model forecasts, which differed in how the state elections were weighted. One model assigned equal weights to all states while the other assigned higher weights to state elections that were held closer to the federal election. The present study used the simple average of both sets of forecasts.
- In addition to their two polling averages, Signal & Rauschen (signalundrauschen.de) also published a model forecast, hereafter referred to as S&R_model. S&R_model is basically a combination of two components, namely the S&R_weighted polling average and an adapted version of the Länder-based model by Kayser and Leininger (2017). The combination of the components followed Küntzler (2017).

As shown in Figure 1.E, the models' MAEs ranged from 2.3 percentage points (S&R_model) to 4.2 percentage points (Chancellor model).

4.4.6 Combined PollyVote forecast

As shown in Figure 2, daily forecasts of the combined PollyVote were calculated using a two-step procedure. In the first step, the individual party-level forecasts were averaged within each of the four component methods:

1. The *combined polls* forecast was calculated as the simple average of the latest party-level estimates of the seven poll aggregators listed in Figure 1.B, which already combined polls from different pollsters (cf. Figure 1.A).
2. The *combined prediction markets* forecast was calculated as the simple average of the daily forecasts of the four available prediction markets.
3. The *combined experts* forecast was the simple average of the forecasts from all experts who participated in a survey round.
4. The *combined models* forecast was the simple average of the daily forecasts from the five available models.

In the second step, the combined PollyVote forecast was calculated by calculating simple (unweighted) averages of the already combined component forecasts.

Figure 3 shows the MAE of the PollyVote and the typical and combined component forecasts. The MAE of the typical component is the error that one would get if one would randomly pick one of the individual forecasts within that component method. The MAE of the combined component is the error that one would get by relying on that method's combined forecast (i.e., after combining within component methods).

The MAE of the PollyVote was 2.4 percentage points, which is less accurate than the combined (2.0 percentage points) and typical (2.1) market forecast, the combined polls (2.3) forecast, and the typical poll aggregator (2.3). Less accurate than the PollyVote were the typical poll (2.4) as well as both combined and typical forecasts from experts and models.

Figure 2: Procedure for calculating the combined PollyVote for forecasting the 2017 German federal election

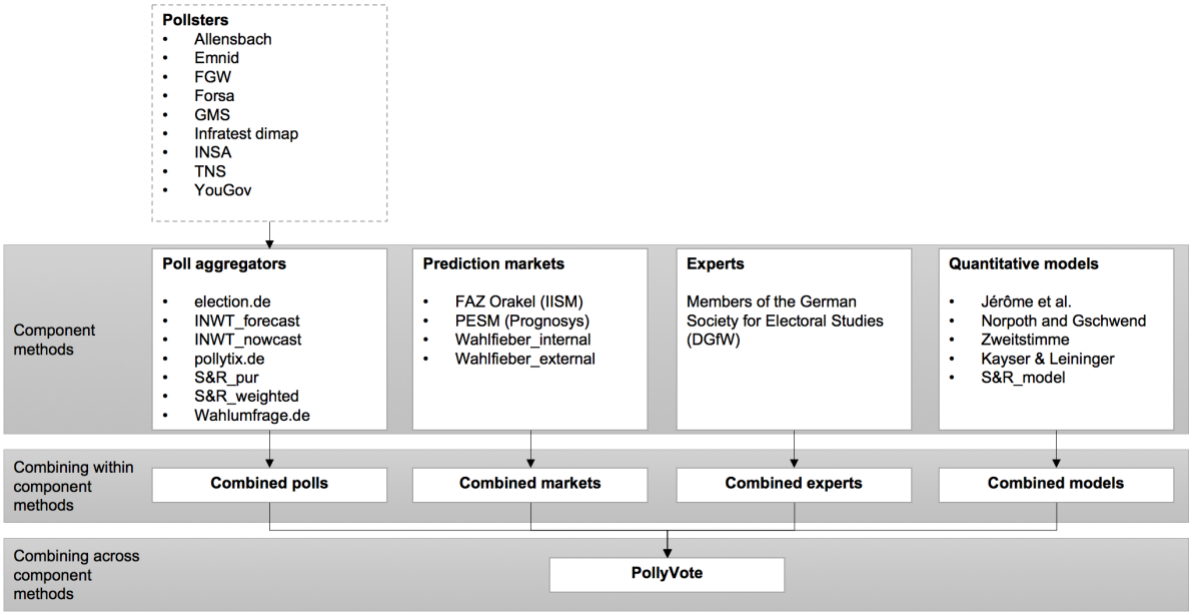
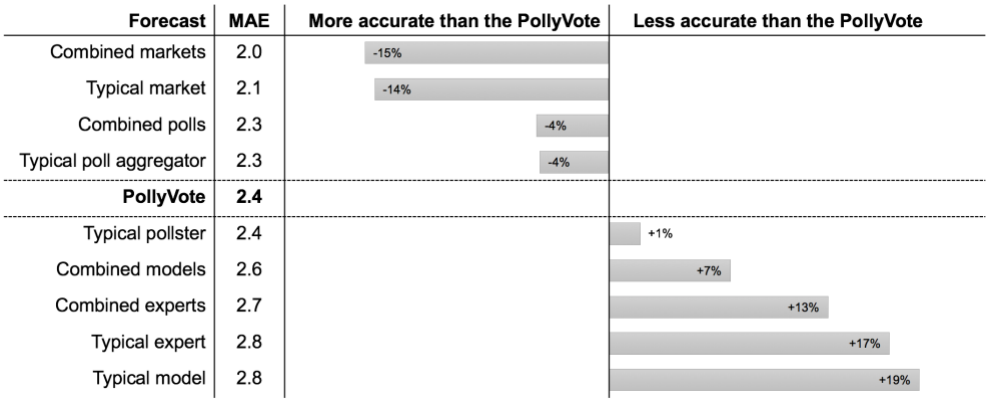


Figure 3: MAE of the PollyVote and the typical and combined component forecasts, calculated across the last 63 days before the 2017 election



5 Forecast accuracy across the 2013 and 2017 elections

The 2017 election was the second time after 2013 that the PollyVote was used to forecast the German federal election. For a detailed analysis of the 2013 results see Graefe (2015). This section reviews the accuracy of the PollyVote and the four component forecasts across both elections in 2013 and 2017.

The PollyVote's general specification was the same in both elections: the PollyVote combined forecasts within and across the following four component methods: polls, prediction markets, expert judgment, and quantitative models (cf., Figure 2). There were, however, differences at the component level. For example, new component forecasts appeared and/or disappeared for poll aggregators, prediction markets, and quantitative models. When it comes to expert judgment, the recruitment of participants changed.

Figure 4 shows the MAE of the PollyVote and 23 other forecasts that were available for both elections. The MAEs were calculated by first averaging the errors across the last 46 days prior to each election, and then averaging the resulting values across both elections. The MAEs ranged from 1.6 percentage points to 3.0 percentage points. With a MAE of 1.8 percentage points, the PollyVote ranked 8th in terms of accuracy. The most accurate individual forecasts were provided by the polling company Infratest dimap, which reduced the error of the PollyVote by 13%. For the second most accurate forecast, the Pollytix poll aggregator, error reduction relative to the PollyVote dropped to 6%. The MAEs of the least accurate forecasts, those from the IISM prediction market and the Chancellor model, were 65% to 67% larger than the corresponding error of the PollyVote.

There was no relationship between a forecast's accuracy in 2013 and 2017 (r=.08). In other words, the performance of an individual forecast in predicting the 2013 election was no indicator for that forecast's performance in predicting the 2017 election.

6 Discussion

The present study applied the PollyVote method of combining forecasts to predicting the 2013 and 2017 German federal elections. The results provide new evidence for the benefits of combining forecasts within and across different methods. In terms of accuracy, the combined PollyVote forecast ranked 8th out of all 24 forecasts that were available for both elections, and was substantially more accurate than the least accurate component forecasts. At the level of individual forecasts, there was no – and in fact an even slightly negative – relationship between past and future accuracy. That is, forecasts that were among the most accurate in 2013 tended to be among the least accurate in 2017. An example are prediction markets, which, on average, provided by far the least accurate forecasts in 2013 but were the most accurate component method in 2017. Similarly, while both 2013 and 2017 were rather good years for German pollsters, it is not guaranteed that the polls will be equally accurate for future elections.

The variance in individual methods' accuracy is a common finding in election forecasting, and of course the very reason why combining is so useful in the first place. This is demonstrated by the PollyVote's performance in predicting the seven U.S. presidential elections from 1992 to 2017. Needless to say, the combined PollyVote did not outperform its most accurate component in each single election (although this has happened). But, over time, as the relative accuracy of individual component forecasts varies, the gains from combining increase. Across the seven elections from 1992 to 2016, the combined PollyVote's MAE of 1.1 percentage points is lower than the corresponding error of any other method (Graefe et al., 2014b). Finally, even if the combined forecast may not be the most accurate forecast in one particular election, and it most likely will not be, combining is still useful in preventing forecasters from picking a poor forecast.

Figure 4: MAE of the PollyVote and other forecasts, 2013 and 2017 elections, calculated across the last 46 days before each election

Forecast	Component method	MAE	More accurate than the PollyVote	Less accurate than the PollyVote
Infratest dimap	Pollster	1.6	-13%	
Pollytix.de	Poll aggregator	1.7	-6%	
Wahlfieber_external	Prediction market	1.7	-5%	
Combined polls	Poll aggregators (combined)	1.7	-5%	
Wahlfieber_internal	Prediction market	1.8	-2%	
Wahlumfrage.de	Poll aggregator	1.8	-1%	
Emnid	Pollster	1.8	0%	
PollyVote		**1.8**		
GMS	Pollster	1.8		0%
Election.de	Poll aggregator	1.9		+1%
Combined models	Quantitative models (combined)	1.9		+2%
Typical pollster	Pollster (typical)	1.9		+2%
Forsa	Pollster	1.9		+5%
INSA	Pollster	1.9		+6%
PESM	Prediction market	1.9		+6%
Combined markets	Prediction markets (combined)	1.9		+6%
Jérôme, Jérôme-Speziari & Lewis-Beck	Quantitative model	2.0		+10%
Typical market	Prediction market (typical)	2.1		+14%
Allensbach	Pollster	2.1		+16%
Combined experts	Experts (combined)	2.1		+17%
Typical model	Quantitative model (typical)	2.2		+20%
Typical expert	Expert (typical)	2.4		+29%
IISM	Prediction market	3.0		+65%
Norpoth & Gschwend	Quantitative model	3.0		+67%

References

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 417-439). New York: Springer.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *OR, 20*(4), 451-468.

Campbell, J. E., Norpoth, H., Abramowitz, A. I., Lewis-Beck, M. S., Tien, C., Erikson, R. S., . . . Cuzán, A. G. (2017). A recap of the 2016 election forecasts. *PS: Political Science & Politics, 50*(2), 331-338.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Cuzán, A. G., Armstrong, J. S., & Jones, R. J. J. (2005). How we Computed the PollyVote. *Foresight: The International Journal of Applied Forecasting, 1*(1), 51-52.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting, 29*(1), 108-121.

Graefe, A. (2014). Accuracy of vote expectation surveys in forecasting elections. *Public Opinion Quarterly, 78*(S1), 204-232. doi:10.1093/poq/nfu008

Graefe, A. (2015). German election forecasting: Comparing and combining methods for 2013. *German Politics, 24*(2), 195-204. doi:10.1080/09644008.2015.1024240

Graefe, A. (2017a). Political Markets. In K. Arzheimer, J. Evans, & M. S. Lewis-Beck (Eds.), *The SAGE Handbook of Electoral Behavior* (Vol. 2, pp. 861-882). London: SAGE.

Graefe, A. (2017b). Prediction market performance in the 2016 U.S. presidential election. *Foresight: The International Journal of Applied Forecasting, 2017*(45), 38-42.

Graefe, A. (2018). Predicting elections: Experts, polls, and fundamentals. *Judgment and Decision Making, 13*(4), 334-344.

Graefe, A., Armstrong, J. S., Jones, R. J. J., & Cuzán, A. G. (2009). Combined Forecasts of the 2008 Election: The PollyVote. *Foresight: The International Journal of Applied Forecasting, 2009*(12), 41-42.

Graefe, A., Armstrong, J. S., Jones, R. J. J., & Cuzán, A. G. (2014a). Accuracy of Combined Forecasts for the 2012 Presidential Election: The PollyVote. *PS: Political Science & Politics, 47*(2), 427-431. doi:doi:10.1017/S1049096514000341

Graefe, A., Armstrong, J. S., Jones, R. J. J., & Cuzán, A. G. (2014b). Combining forecasts: An application to elections. *International Journal of Forecasting, 30*(1), 43-54.

Graefe, A., Armstrong, J. S., Jones, R. J. J., & Cuzán, A. G. (2017). Assessing the 2016 U.S. presidential election popular vote forecasts. In A. Cavari, R. Powell, & K. Mayer (Eds.), *The 2016 Presidential Election: The Causes and Consequences of a Political Earthquake* (pp. 137-158). Lanham, MD: Lexington Books.

Graefe, A., Küchenhoff, H., Stierle, V., & Riedl, B. (2015). Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems. *International Journal of Forecasting, 31*(3), 943-951.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231-237.

Hillygus, D. S. (2011). The Evolution of Election Polling in the United States. *Public Opinion Quarterly, 75*(5), 962-981. doi:10.1093/poq/nfr054

Jérôme, B., Jérôme-Speziari, V., & Lewis-Beck, M. S. (2017). The Grand Coalition reappointed but Angela Merkel on borrowed time. *PS: Political Science & Politics, 50*(3), 683-685.

Jones, R. J. J., & Cuzán, A. G. (2013). *Expert Judgment in Forecasting American Presidential Elections: A Preliminary Evaluation*. Paper presented at the Annual Meeting of the American Political Science Association (APSA), Chicago.

Kayser, M., & Leininger, A. (2017). A Länder-based forecast of the 2017 German Bundestag election. *PS: Political Science & Politics, 50*(3), 689-692.

Kernell, S. (2000). Life before polls: Ohio politicians predict the 1828 presidential vote. *PS: Political Science & Politics, 33*(3), 569-574.

Küntzler, T. (2017). Using Data Combination of Fundamental Variable-Based Forecasts and Poll-Based Forecasts to Predict the 2013 German Election. *German Politics (Online First)*, 1-19. doi:10.1080/09644008.2017.1280781

Lewis-Beck, M. S. (2005). Election forecasting: principles and practice. *The British Journal of Politics & International Relations, 7*(2), 145-164.

Linzer, D. A. (2013). Dynamic Bayesian Forecasting of Presidential Elections in the States. *Journal of the American Statistical Association, 108*(501), 124-134. doi:10.1080/01621459.2012.737735

Munzert, S., Stötzer, L., Gschwend, T., Neunhoeffer, M., & Sternberg, S. (2017). Zweitstimme.org. Ein strukturell-dynamisches Vorhersagemodell für Bundestagswahlen. *PVS Politische Vierteljahresschrift, 58*(3), 418-441. doi:10.5771/0032-3470-2017-3-418

Norpoth, H., & Gschwend, T. (2017). Chancellor model predicts a change of the guards. *PS: Political Science & Politics, 50*(3), 686-688.

Pasek, J. (2015). Predicting elections: Considering tools to pool the polls. *Public Opinion Quarterly, 79*(2), 594-619. doi:10.1093/poq/nfu060

Rhode, P. W., & Strumpf, K. S. (2004). Historical presidential betting markets. *Journal of Economic Perspectives, 18*(2), 127-141.

Rhode, P. W., & Strumpf, K. S. (2014). The long history of political betting markets: An international perspective. In L. Vaughan Williams & D. S. Siegel (Eds.), *The Oxford Handbook of the Economics of Gambling* (pp. 560-586). Oxford: Oxford University Press.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780-805.

Stegmaier, M., & Norpoth, H. (2017). Election Forecasting. In L. S. Maisel (Ed.), *Oxford Bibliographies in Political Sciences*: Oxford University Press.