International Journal of Forecasting 39 (2023) 170–177

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Embrace the differences: Revisiting the PollyVote method of combining forecasts for U.S. presidential elections (2004 to 2020)

Andreas Graefe

Macromedia University of Applied Sciences, Munich, Germany

ARTICLE INFO

Keywords:
Combining forecasts
Evaluating forecasts
Monitoring forecasts
Political forecasting
Election forecasting

ABSTRACT

While combining forecasts is well-known to reduce error, the question of how to best combine forecasts remains. Prior research suggests that combining is most beneficial when relying on diverse forecasts that incorporate different information. Here, I provide evidence in support of this hypothesis by analyzing data from the PollyVote project, which has published combined forecasts of the popular vote in U.S. presidential elections since 2004. Prior to the 2020 election, the PollyVote revised its original method of combining forecasts by, first, restructuring individual forecasts based on their underlying information and, second, adding naïve forecasts as a new component method. On average across the last 100 days prior to the five elections from 2004 to 2020, the revised PollyVote reduced the error of the original specification by eight percent and, with a mean absolute error (MAE) of 0.8 percentage points, was more accurate than any of its component forecasts. The results suggest that, when deciding about which forecasts to include in the combination, forecasters should be more concerned about the component forecasts' diversity than their historical accuracy.

1. Introduction

In 2004, the PollyVote launched as a long-term project to demonstrate the value of evidence-based forecasting to a broad audience by applying forecasting principles to the high-profile application of forecasting elections (Graefe, 2019; Graefe et al., 2014).[1] At its core, the PollyVote builds on the principle of combining forecasts, a major finding from half a century of forecasting research, tracing back to Bates and Granger (1969). Given that combining forecasts is simple and effective, it comes as no surprise that the principle has long been successfully applied in various domains such as economics, meteorology, or sports, to name only a few (Clemen, 1989).

There are several benefits of combining forecasts. First, combining forecasts reduces bias. Any single forecast cannot capture all important information because the underlying method is limited in the amount of information it can include. For example, regression-based models are limited in the number of variables they can incorporate, particularly if historical data for model estimation are few (Armstrong, 2012), as is the case with many practical problems. As a result, the forecast may be biased due to omitted information. If one relies on forecasts that use different methods and different information, one lowers the risk of omitting relevant information and likely reduces bias.

Second, combining forecasts protects from picking a poor forecast. This is important, because people tend to be overconfident, and – wrongly – believe that they can identify the best forecast among a set of forecasts (Soll & Larrick, 2009). For example, people may think that the best forecasts are those that did well in the past. However,

E-mail address: graefe.andreas@gmail.com.

[1] PollyVote was founded in 2004 by J. Scott Armstrong, Alfred G. Cuzán, and Randall J. Jones Jr.

https://doi.org/10.1016/j.ijforecast.2021.09.010

the relative accuracy of different methods tends to vary across time and events, and past accuracy can be a poor predictor of future accuracy (Graefe et al., 2015). Combining forecasts avoids the danger of picking a poor forecast. The simple average of all available forecasts will always be at least as accurate as a randomly chosen forecast. Note that this is different from saying that the combined forecast yields only average performance, which is a common misperception (Larrick & Soll, 2006).

Rather, third, the opposite is true. Given that the accuracy of individual forecasts varies across time, in the long run, the combined forecast will be more accurate than a randomly chosen forecast, and may even outperform the most accurate individual forecast. Graefe et al. (2014) analyzed the PollyVote's accuracy in predicting the popular two-party vote across the U.S. presidential elections from 1992 to 2012 (1992–2000 as ex post forecasts). The PollyVote did not provide the most accurate forecasts in each single election. Yet, on average across the six elections, the PollyVote was more accurate than forecasts from four benchmark methods.

Research has long focused on how to best combine forecasts, for example, by trying to find optimal schemes for weighting the individual forecasts. It is, however, difficult to find evidence in support of such a strategy, and available evidence suggests that the weighting of forecasts is uncritical. Often, a simple forecast average outperforms more complex weighting schemes (Genre et al., 2013; Graefe et al., 2015; Stock & Watson, 2004). In their seminal work, Bates and Granger (1969) pointed out that combining forecasts is most beneficial if the errors of individual forecasts are negatively correlated, as the errors would then cancel each other out in the aggregate. Armstrong (2001) suggested that one may be able to create such conditions by combining forecasts that use different data and/or methods. For practical problems, however, forecast errors often correlate positively, as forecasters tend to rely on similar methods and data. For example, when it comes to election forecasting, most forecasters use – often already correlated – data on public opinion or economic fundamentals, or combinations thereof. Recent research on combining has thus focused on helping forecasters to identify a diverse set of forecasts, for example, by analyzing individual forecasts' past performance and coherence (Thomson et al., 2019).

The present paper discusses the 2020 PollyVote specification as an alternative strategy to combining diverse forecasts. Rather than selecting certain subsets of forecasts based on their statistical properties, the PollyVote categorizes forecasts based on prior domain knowledge about how the forecasts differ in their underlying information. The forecast accuracy of this approach is compared with the PollyVote's previous specification as well as to various benchmark forecasts.

2. The 2020 PollyVote

Since 2004, the PollyVote has averaged forecasts within and across different component methods, each of which has been validated for forecasting election outcomes. For forecasting the U.S. presidential elections in 2004 and 2008, as well as the German federal elections in 2013 and 2017 (Graefe, 2019), the PollyVote averaged forecasts within and across four component methods, namely, polls, betting – also known as prediction – markets, expert judgment, and econometric models. In the U.S. case, this original specification changed over time, motivated by efforts to implement advances from forecasting research. In particular, two new component methods were added, namely, index models prior to the 2012 election, and citizen forecasts prior to 2016. That is, the PollyVote averaged forecasts within and across six different component methods to predict the popular two-party vote in the 2016 U.S. presidential election.

2.1. Revisions

After the 2016 election, the PollyVote method underwent two important changes, which were motivated by the project's mission to incorporate evidence-based findings from forecasting research as well as to improve the communication of forecasts.

1. Naïve forecasts were added as a new component method, guided by the recommendation to better acknowledge forecast uncertainty by being conservative (Armstrong et al., 2015).
2. Following the advice to combine forecasts that differ in their underlying information (Armstrong, 2001), expectations- and model-based forecasts were restructured to better reflect such differences.

This resulted in the creation of a new method component named *expectations*, which encompasses the three formerly separate component methods betting markets, expert judgment, and citizen forecasts. Forecasts from these methods are similar in that they all rely on people's expectations of what will happen on Election Day – they only differ with respect to who the participants are, and how their expectations are aggregated into a forecast. The rationale behind the decision to collapse these methods into an own category – rather than keeping them as separate components – was to avoid bias due to overweighting expectation-based forecasts.

With respect to models-based forecasts, two formerly separate components (i.e., econometric and index models) were collapsed into a single component. The idea was to move away from differentiating models based on the underlying statistical method (i.e., how to weigh the variables in a model) but rather based on the information the models incorporate. This led to distinguishing models that are purely retrospective, purely prospective, and those that are mixed in drawing both on retrospective and prospective voting theories.

These changes to the PollyVote structure were implemented prior to forecasting the 2020 U.S. presidential election (Armstrong & Graefe, 2021). That is, no data from the 2020 election were used to develop the revised specification. In fact, no historical data were used at all, as the changes were motivated by findings from prior research. After deciding on which changes to implement, an ex-post analysis was conducted to test how the revised version would have performed when forecasting the four

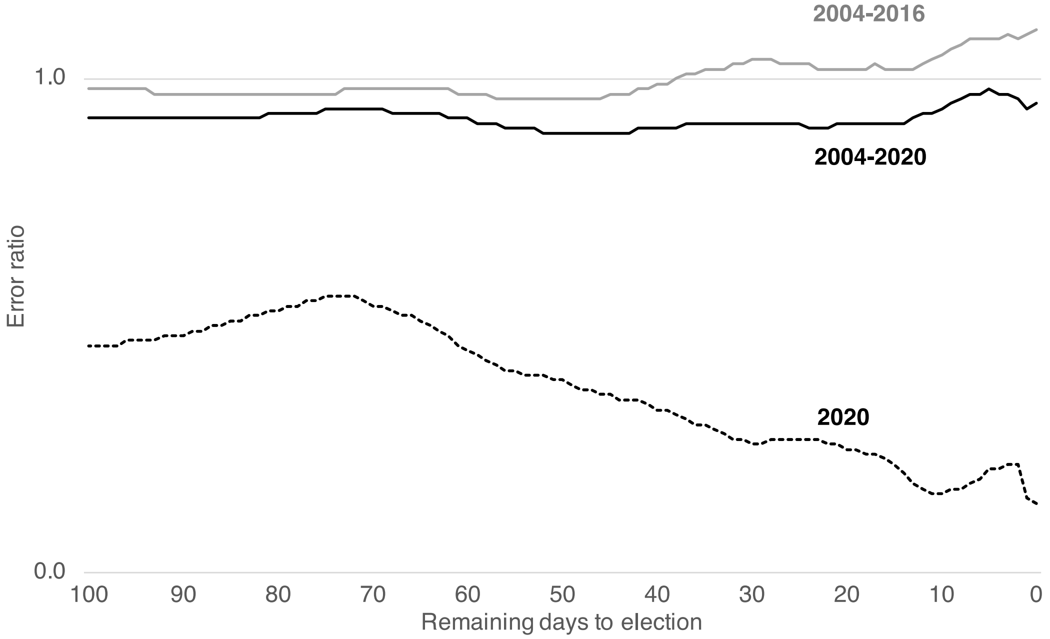

Fig. 1. Error ratio of the revised vs. the original PollyVote specification. For each day, the chart shows the error ratio of forecasts from the revised 2020 PollyVote specification (developed prior to the 2020 election) vs. the PollyVote specification used in 2016, from that day forward until Election Day. Values below 1 mean that the revised PollyVote specification was more accurate, values above 1 mean that the 2016 PollyVote specification was more accurate. An error ratio of 1 means no difference in accuracy. For example, across the last 100 days across the elections from 2004 to 2020 (black solid line), the error ratio was 0.92. That is, the error of the revised PollyVote specification was 8% (1–0.92) lower than the corresponding error of the original PollyVote specification.

elections from 2004 to 2016. The results showed that, across the last 100 days before each election, the revised 2020 PollyVote specification would have reduced the error of the 2016 specification on average by merely 2%. For shorter forecast horizons, starting at around 40 days prior to the election, the revised PollyVote would have been slightly less accurate than the 2016 specification (cf. Fig. 1, grey line). Despite these results, it was decided to remain with the revised specification. The reason was that the analysis is only based on four elections, whereas the suggested revisions were well-grounded in evidence-based forecasting research. Thus, gains in accuracy from using the revised approach were expected in the long run. Furthermore, forecast accuracy aside, the revisions improved the underlying logic of the PollyVote's structure and reduced complexity (four main components instead of six). Hence, the 2020 specification was expected enhance users' understanding of the PollyVote and its components.

2.2. Structure

Table 1 shows the structure for the 2020 PollyVote. The PollyVote forecast is the average forecast calculated across four main component methods (ordered by their average forecast accuracy across the last 100 days prior to the five U.S. presidential elections from 2004 to 2020): (I) models, (II) expectations, (III) polls, and (IV) naïve forecasts. This section explains each component method and the underlying forecasts.

I. Models

Available models for forecasting the popular vote in U.S. presidential elections rely on aggregate data. Thereby, the choice of predictor variables depends on the models' underlying theories of voting, which can be retrospective, prospective, or a combination of both. The PollyVote uses this distinction to group the various models. Note that model specifications may also differ with respect to assumptions regarding sociotropic voting and/or pocketbook voting. Whereas sociotropic voting assumes that voters are altruistic and evaluate the incumbent government based on perceived national (economic) conditions, pocketbook voting assumes that voters prefer candidates or parties under which they expect to be better off personally (Elinder et al., 2015). However, given that virtually all available forecasting models assume sociotropic voting, this distinction cannot be used for model classification.

I.A. Retrospective models. Retrospective models assume that voters reward incumbents for good performance and punish them otherwise. In other words, elections are considered referenda on the incumbent government's past (economic) performance. When it comes to the choice of economic predictor variables, all available retrospective models essentially assume sociotropic voting in employing structural data about national economic (or political) conditions.

I.A.1. Fundamentals-only models Some forecasting models measure performance solely through national economic (e.g., gross domestic product (GDP), unemployment, inflation, etc.) or political (e.g., fiscal spending, war fatalities) variables, the so-called fundamentals. Given that these models completely ignore public opinion variables, they are hereafter referred to as *fundamentals-only* models. Such models have become rare. For the 2020 election, only forecasts from the classic model by Fair (2009) were available. Similar models, which used to be around in earlier elections, such as the Bread and Peace model (Hibbs, 2000) or the Fiscal model (Cuzán, 2012), did not publish forecasts in 2020. The likely reason for the disappearance of fundamentals-only models from the forecasting scene is that, in terms of forecast accuracy, these models cannot compete with models that incorporate more information. Yet, their disappearance is misfortunate. First, fundamentals-only models have explanatory power in trying to estimate how certain fundamentals affect aggregated vote choice. Second, fundamentals-only models provide useful indicators for the direction of polling error (Graefe, 2018), and thus should help to improve the accuracy of a combined forecast.

I.A.2. Fundamentals-plus models The other category of purely retrospective models consists of those that additionally also incorporate public sentiment on the incumbent president's job performance. For forecasting the 2020 election, only the forecast from Lewis-Beck and Tien (2021) was available in this category, as Abramowitz (2021) decided against publishing a forecast using his classic Time-for-Change model due to the specific circumstances with respect to the Coronavirus pandemic. These models, hereafter referred to as *fundamentals-plus*, are superior to fundamentals-only models in terms of forecast accuracy. It should however be noted that the explanatory value of fundamentals-plus models is small, because using the proxy variable of presidential job approval makes it impossible to differentiate effects of economic and noneconomic factors as well as candidate characteristics on the election outcome.

I.B. Prospective models. Prospective models assume that voters look ahead and evaluate the candidates based on their future promises, and vote for the candidate with whom they expect to be better off – either themselves or the country as a whole (Hsieh et al., 1998). The 2020 PollyVote included two models that fall in that category, both of which use aggregate data from public opinion polls. One is the Issues and Leaders model, which predicts the election outcome based on how voters perceive the candidates' leadership and issue-handling skills (Graefe, 2021). The one other is the Big-Issue model, which uses information about which candidate voters think will do a better job in handling the most important problem facing the country (Graefe & Armstrong, 2012).

I.C. Mixed models. Mixed models are neither purely retrospective nor purely prospective, but incorporate both types of information. For example, any model that uses fundamental economic data as well as trial-heat polls would be classified as a mixed model. This category best reflects contemporary election forecasting models, and includes classic models developed by academics (Erikson & Wlezien, 2021; Lichtman, 2008) as well as widely popular models such as those published by *The Economist* and *FiveThirtyEight*. When it comes to the choice of economic predictor variables, virtually all available models use structural data about national economic conditions. That is, election forecasting models essentially assume sociotropic voting. A notable exception used to be Holbrook (2016), who combined 'aggregate satisfaction with personal finances' and 'presidential approval' to calculate an 'index of national conditions', but did not publish a forecast from that model after the 2016 election. Forecast accuracy aside, the explanatory power of mixed models is – similar to fundamentals-plus models – limited, due to the confounding effects of using both economic fundamentals and public opinion polls in the same model.

II. Expectations

II.A. Expert judgment. When predicting the future, likely no method has a longer history than asking domain experts about what is going to happen. When it comes to election forecasting, one may expect experts to be able to improve upon the accuracy of polls, for example, by putting polls into historical perspective or accounting for any variance due to campaign events. Surprisingly, there is little evidence on the relative performance of polls and expert judgment. One study compared the forecast accuracy of a polling average to 452 individual expert forecasts made across the four U.S. presidential elections from 2004 to 2016. The results showed that 62% of experts' forecasts correctly predicted the directional error of polls. However, the typical expert's error was 7% higher than the corresponding error of a polling average (Graefe, 2018).

As with all forecasts, combining individual expert judgment should improve accuracy. Between April 2020 and Election Day, I asked 15 political science professors from various U.S. universities to forecast the candidates' national vote shares, first once a month, and more frequently closer to the election. Their mean forecast was used as the PollyVote's expert forecast.

II.B. Betting markets. Betting (or prediction) markets allow participants to bet money on the election outcome. Participants trade contracts for each candidate, whereby a contract's price reflects a candidate's predicted vote share. Participants self-select and have an incentive to be right, as they win (or lose) money depending on the accuracy of their predictions.

Until the 2020 U.S. presidential election, there was only the small-scale Iowa Electronic Markets (IEM), operated by the University of Iowa since 1988, that provided vote share forecasts (Gruca & Rietz, 2021). This market lacks efficiency due to its low volume and trading restrictions (i.e., participants cannot invest more than $500). In analyzing forecast accuracy across the last 96 days prior to each of the four elections from 2004 and 2016, Graefe (2017) found that the IEM's error was 38% higher than the corresponding error of expert judgment and 57% higher than the error of the RealClearPolitics poll average.

In 2020, for the first time, PollyVote also incorporated forecasts from another betting market called PredictIt, operated by Victoria University of Wellington in New

Table 1
Forecast error of the PollyVote and its components (2004–2020).

				Final 2020 forecast	Error (%-points)	MAE across last 100 days to election (%-points)						
						2004–2020	2020	2016	2012	2008	2004	Source
PollyVote				52.2	−0.1	0.8	0.2	0.6	0.6	2.0	0.6	(Armstrong & Graefe, 2021)
I			Models	**52.9**	**0.6**	**1.0**	**0.7**	**0.3**	**0.3**	**0.9**	**2.7**	
	A		***Retrospective***	***52.3***	***0.0***	***2.0***	***1.6***	***2.3***	***2.2***	***1.2***	***2.8***	
		1	Fundamentals-only	47.9	−4.4	2.7	2.0	3.4	2.3	1.9	3.9	
			Berry & Bickers	na	–	na	na	na	nya	na	na	(Berry & Bickers, 2012)
			Bread and Peace	na	–	na	na	nya	4.5	1.8	nya	(Hibbs, 2000)
			Fair	47.9	−4.4	4.0	2.0	7.1	2.5	2.1	6.3	(Fair, 2009)
			Fiscal	na	–	na	na	nya	5.7	nya	nya	(Cuzán, 2012)
			Haynes & Stone	na	–	na	na	na	na	nya	na	(Haynes & Stone, 2008)
			Jerome & Jerome	na	–	na	na	1.0	0.5	0.4	na	(Jerôme & Jerôme-Speziari, 2016)
			Primary	na	–	na	na	3.6	1.2	3.6	3.5	(Norpoth, 2016)
			Proxy	na	–	na	na	na	0.8	na	na	(Lewis-Beck & Tien, 2012)
		2	Fundamentals-plus	56.7	4.4	na	nya	1.1	2.0	nya	nya	
			Jobs	na	–	na	na	na	nya	nya	nya	(Lewis-Beck & Tien, 2012)
			Lewis-Beck & Tien	56.7	4.4	na	nya	nya	na	na	na	(Lewis-Beck & Tien, 2021)
			Klarner	na	–	na	na	na	0.7	nya	na	(Klarner, 2012)
			Time-for-change	na	–	na	na	2.5	1.4	nya	nya	(Abramowitz, 2021)
	B		***Prospective***	***53.5***	***1.2***	***na***	***0.5***	***3.3***	***2.2***	***1.3***	***na***	
			Big-Issue	52.8	0.5	na	0.4	0.2	0.8	na	na	(Graefe & Armstrong, 2012)
			Bio-index	na	–	na	na	7.6	3.1	na	na	(Armstrong & Graefe, 2011)
			Issue-index	na	–	na	na	4.8	6.5	1.3	na	(Graefe & Armstrong, 2013)
			Issues and Leaders	54.1	1.8	na	0.8	0.7	1.1	na	na	(Graefe, 2021)
	C		***Mixed***	***52.9***	***0.6***	***0.9***	***0.5***	***0.2***	***0.5***	***0.5***	***2.6***	
			538 (polls-plus)	54.1	1.8	na	1.3	0.9	0.6	na	na	fivethirtyeight.com
			Convention bump	na	–	na	na	nya	nya	na	na	(Campbell, 2016)
			Crosstab	na	–	na	na	nya	na	na	na	thecrosstab.com
			DeSart	54.8	2.5	na	2.4	1.8	na	na	na	(DeSart, 2021)
			DeSart & Holbrook	54.4	2.1	na	nya	nya	nya	nya	na	(DeSart & Holbrook, 2003)
			Economist	54.2	1.9	na	2.5	na	na	na	na	economist.com
			Holbrook	na	–	na	na	nya	nya	nya	nya	(Holbrook, 2016)
			Lichtman's Keys	52.1	−0.2	na	nya	0.9	3.0	0.3	2.2	(Lichtman, 2008)
			Lockerbie	44.8	−7.4	na	nya	nya	1.8	4.5	nya	(Lockerbie, 2021)
			Trial-heat	na	–	na	na	nya	nya	nya	nya	(Campbell, 2016)
			Vox.com	na	–	na	na	nya	na	na	na	vox.com/a/trump-tax
			Wlezien & Erikson	55.7	3.4	na	nya	0.9	nya	nya	nya	(Erikson & Wlezien, 2021)
II			**Expectations**	**52.5**	**0.2**	**1.1**	**0.5**	**2.7**	**0.5**	**1.3**	**0.7**	
	A		***Experts***	***53.4***	***1.1***	***na***	***1.4***	***2.1***	***1.2***	***1.7***	***nya***	(Armstrong & Graefe, 2021)
	B		***Betting markets***	***53.2***	***0.9***	***1.9***	***1.4***	***4.7***	***1.5***	***1.2***	***0.6***	
			Predictit	*52.2*	*−0.1*	*na*	*0.3*	*na*	*na*	*na*	*na*	predictit.org
			IEM	*54.2*	*1.9*	*2.2*	*2.7*	*5.0*	*1.5*	*1.2*	*0.7*	(Gruca & Rietz, 2021)
	C		***Citizen forecasts***	***50.8***	*−1.5*	***1.2***	***1.7***	***1.1***	***0.4***	***1.7***	***1.1***	(Graefe, 2014)
III			**Polls**	**54.0**	**1.8**	**1.5**	**1.8**	**1.6**	**1.1**	**1.7**	**1.4**	
			270 to win	na	–	na	na	2.2	na	na	na	270towin.com
			538 (polls-only)	54.4	2.1	na	2.2	1.7	na	na	na	fivethirtyeight.com
			Economist polls	54.2	1.9	na	2.1	na	na	na	na	economist.com
			Election Projection	na	–	na	na	1.6	na	na	na	electionprojection.com
			HuffPost Pollster	na	–	na	na	2.2	na	na	na	elections.huffingtonpost.com/pollster
			PEC	53.8	1.5	na	1.2	0.9	na	na	na	election.princeton.edu
			RealClearPolitics	53.8	1.5	1.5	1.8	1.5	1.1	1.7	1.4	realclearpolitics.com
			TPM Poll Tracker	na	–	na	na	1.7	na	na	na	talkingpointsmemo.com
IV			**Naïve**	**49.4**	**−2.9**	**2.2**	**2.9**	**2.0**	**1.2**	**4.5**	**0.4**	
			Electoral-cycle	48.8	−3.5	2.8	3.5	2.9	0.5	5.4	2.0	(Norpoth, 2014)
			50/50	50.0	−2.3	2.1	2.3	1.1	2.0	3.7	1.2	(Armstrong & Graefe, 2021)

Final 2020 forecasts refer to the Democratic share of the two-party vote; few forecasts were available for all five elections from 2004 to 2020 and, of those that were, not all were available across the last 100 days before the election (na: forecast not available for that election; nya: forecast available, but published later than 100 days before that election). Use online tool available at https://tinyurl.com/PollyVote-IJF to analyze results for different forecast horizons.

Zealand. PredictIt offered 16 contracts on what will be the final popular vote margin, and their prices were converted into vote share forecasts for Trump and Biden. The forecast of the PollyVote's betting market component was the simple average of the daily IEM and PredictIt forecasts.

II.C. Citizen forecasts. Citizen forecasts are derived from survey respondents' answers to the vote expectation question, which asks them who they think will win the election (in addition to who they would vote for). As shown by Graefe (2014), the aggregate answers to that question can be translated into highly accuracy vote share forecasts using the incumbent vote share as the dependent variable in a simple linear regression. Across the last 100 days prior to the seven presidential elections from 1988 to 2012, citizen forecasts were more accurate than polls, betting markets, models, and expert forecasts.

Given the method's predictive accuracy, it is unfortunate that the vote expectation question is rarely included in public opinion surveys. Between April 2020 and Election Day, only 38 surveys were found that asked the question, of which 35 were conducted by YouGov and one each by Gallup, Fox News, and Monmouth. For each day, the PollyVote's citizen forecast uses the most recent result from a vote expectation question, translated into a two-party vote share forecast using the equation originally estimated by Graefe (2014), but updated including 2016 data.

III. Polls

Technically, polls that ask respondents for which candidate they are going to vote do not provide forecasts. Polls merely capture respondents' vote preferences at a particular point in time, which may change until the election. Therefore, polls tend to be less accurate the farther away the election. Also, results of polls conducted around the same time often vary wildly, which can be partly explained by different methodologies used across pollsters (Erikson & Wlezien, 2012). The good news is that many of the biases associated with individual polls tend to cancel out when calculating polling averages, which have become increasingly popular. For the 2020 election, the PollyVote's polls component averaged the daily estimates from four polling aggregators (i.e., FiveThirtyEight, Economist, Princeton Election Consortium, and RealClearPolitics), which slightly differ in their methodology (e.g., with respect to which polls they include and how they weigh them).

Polling averages cannot, however, account for systematic polling error, for example, due to nonresponse (Gelman et al., 2016). In such situations, polls tend to err in the same direction, and the value of combining polls is limited. This is what happened in both 2016 and 2020, when most individual polls, and hence also polling averages, substantially underestimated Donald Trump's vote share.

IV. Naïve forecasts

Complexity tends to harm forecast accuracy. That is, the accuracy of very simple models, such as a naïve no-change model, is often difficult to beat by more complex models (Green & Armstrong, 2015). Reasons for using a no-change model could be that one expects that the situation will not change, or that one cannot predict the direction of change. Put differently, the rationale behind incorporating a naïve component is to acknowledge a situation's underlying uncertainty and thereby adhere to the principle of conservatism in forecasting (Armstrong et al., 2015). In addition, naïve forecasts are unlikely to correlate with other forecasts and should thus contribute to the accuracy of a combined forecast.

To forecast the 2020 election, naïve forecasts were added as a new component to the PollyVote, which consisted of the average of two naïve models, namely, (A) the electoral cycle (Norpoth, 2014) and (B) a 50/50 model. The electoral cycle model uses the incumbent vote of the two most recent elections as predictor variables in a linear model estimated based on data from all elections since 1828. The 50/50 model assumes that both major-party candidates will gain 50% of the popular vote and thus represents the age of political polarization.

3. Forecast accuracy

3.1. 2020 election

The 2020 PollyVote's popular two-party vote forecast was remarkable with respect to both stability and accuracy. Since its launch on May 15, the PollyVote consistently – and correctly – predicted that Joe Biden would win the popular vote. Across the 172 days to Election Day, Biden's predicted vote share averaged 52.2 percent and remained within 51.5 and 52.7 percent, a narrow range of only 1.2 percentage points (standard deviation: 0.3 points). As shown in Table 1, the PollyVote's final forecast (52.2% for Biden, thereby matching the average forecast since its first release) underestimated Biden's final vote share by merely 0.1 percentage points: Biden eventually won 52.3 percent of the two-party vote (vs. 47.7 percent for Trump). The PollyVote' Election Eve forecast was more accurate than the forecasts from each of its four component methods, three of which overpredicted Biden's vote share (expectations by 0.2 points, models by 0.6 points, and polls by 1.7 points), while naïve models underpredicted it. Of all individual component forecasts, only one (the PredictIt betting market) matched the accuracy of the PollyVote at Election Eve, while retrospective models (i.e., the combination of fundamentals-only and fundamentals-plus models) were even more accurate in hitting the final election outcome.

Apart from providing entertainment and informing last minute voters, Election Eve forecasts are of limited practical value. Hence, when evaluating forecast accuracy, decision-makers should focus on longer forecast horizons. Table 1 also shows the mean absolute error (MAE) calculated across forecasts made for the last 100 days until the election. With an MAE of 0.2 points, the PollyVote was more accurate than each of its four component methods, and any of its individual component forecasts.

3.2. 2004 to 2020 elections

Although the PollyVote did exceptionally well in 2020, its performance is in line with previous elections. Table 1 shows the MAE of the PollyVote and its components across the last 100 days for each of – and across – the elections from 2004 to 2020. With an MAE of 0.8 percentage points across the five elections, the PollyVote was more accurate than each of its component methods, with error reductions of 15% compared with models (MAE: 1.0), 28% compared with expectations (MAE: 1.1), 47% compared with polls (MAE: 1.5), and 63% compared with naïve forecasts (MAE: 2.2). Compared with the best individual forecasting method (citizen forecasts; MAE: 1.2), the PollyVote reduced error by 33%.

4. Discussion

The PollyVote demonstrated once again that combining forecasts is an effective means to generating accurate forecasts. In 2020, and on average across the last 100 days prior to the five elections from 2004 to 2020, the combined PollyVote was more accurate than any of its component forecasts.

The revisions made to the PollyVote method prior to the 2020 election (i.e., restructuring expectations- and models-based forecasts, and adding naïve forecasts as a new component) yielded large gains in accuracy. When predicting the 2020 election, the revised PollyVote specification reduced the error of the 2016 specification by 45% or more, depending on the forecast horizon (cf. Fig. 1, dotted black line). Across all five elections from 2004 to 2020, error reduction across the last 100 days from using the revised version was 8% – a considerable increase in accuracy, given that the error of the PollyVote was already quite low (cf. Fig. 1, solid black line).

These accuracy gains were achieved despite adding naïve forecasts as a new component, which are clearly less accurate than all other components, both in 2020 and historically. This result may appear counterintuitive. After all, one could think that inaccurate forecasts should be excluded from a combined forecast. However, combining can be beneficial even when adding a forecast that is known to be less accurate than other forecasts. Herzog and Hertwig (2009) illustrated this for the simple case of averaging two forecasts, showing that the error of the less accurate forecast can be up to three times the error of the more accurate forecast – and combining both forecasts would still improve accuracy, as long as both individual forecasts bracket the true value. In other words, when deciding which forecasts to include in the combination, forecasters should be more concerned about the component forecasts' diversity than their historical accuracy.

The PollyVote does not use historical accuracy as a criterion when deciding which forecasts to include. As shown in Table 1, many forecasts have been included over time, but few were available for all five elections since 2004. Of those that were, even fewer provided forecasts for longer forecast horizons (e.g., 100 days prior to Election Day). This, along with the fact that forecasts often disappear from the public domain after an event has taken place, makes accuracy comparisons of different forecasts challenging if not impossible. To lower that barrier, I created an online tool based on the PollyVote's unique collection of historical forecasts, which enables users to compare different forecasts' relative accuracy over time. The tool is available at: https://tinyurl.com/PollyVote-IJF.

Apart from improving forecast accuracy, the revised PollyVote structure aims to help people's understanding about what happens underneath the combined forecast. In particular, the revised PollyVote uncovers differences in forecasts depending on their underlying information, for example, whether a model is based on retrospective and/or prospective voting theories. The new structure also reveals which areas of forecasting are saturated (e.g., the PollyVote incorporated forecasts from seven mixed models in 2020, with little variation except for one outlier) or under-researched (e.g., not a single model published in 2020 assumed pocketbook voting). In doing so, the PollyVote's systematic collection and categorization of forecasts can help assess how the field of election forecasting develops over time. For example, Table 1 shows that, over the past decade, fundamentals-only models have given way to mixed models that also include polling data, such as those published by *FiveThirtyEight.com* and *The Economist*.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abramowitz, A. I. (2021). It's the pandemic, stupid! a simplified model for forecasting the 2020 presidential election. *PS: Political Science & Politics*, *54*(1), 52–54. http://dx.doi.org/10.1017/S1049096520001389.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*, (pp. 417–439). Springer US, http://dx.doi.org/10.1007/978-0-306-47630-3_19.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting*, *28*(3), 689–694. http://dx.doi.org/10.1016/j.ijforecast.2012.02.001.

Armstrong, J. S., & Graefe, A. (2011). Predicting elections from biographical information about candidates: A test of the index method. *Journal of Business Research*, *64*(7), 699–706. http://dx.doi.org/10.1016/j.jbusres.2010.08.005.

Armstrong, J. S., & Graefe, A. (2021). The PollyVote popular vote forecast for the 2020 US presidential election. *PS: Political Science & Politics*, *54*(1), 96–98. http://dx.doi.org/10.1017/S1049096520001420.

Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, *68*(8), 1717–1731.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468. http://dx.doi.org/10.1057/jors.1969.103.

Berry, M. J., & Bickers, K. N. (2012). Forecasting the 2012 presidential election with state-level economic indicators. *PS: Political Science & Politics*, *45*(4), 669–674. http://dx.doi.org/10.1017/S1049096512000984.

Campbell, J. E. (2016). The trial-heat and seats-in-trouble forecasts of the 2016 presidential and congressional elections. *PS: Political Science & Politics*, *49*(4), 664–668. http://dx.doi.org/10.1017/S104909651600127X.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. http://dx.doi.org/10.1016/0169-2070(89)90012-5.

Cuzán, A. G. (2012). Forecasting the 2012 presidential election with the fiscal model. *PS: Political Science & Politics*, *45*(4), 648–650.

DeSart, J. A. (2021). A long-range state-level forecast of the 2020 presidential election. *PS: Political Science & Politics*, *54*(1), 73–76. http://dx.doi.org/10.1017/S1049096520001468.

DeSart, J. A., & Holbrook, T. M. (2003). Statewide trial-heat polls and the 2000 presidential election: A forecast model. *Social Science Quarterly*, *84*(3), 561–573, http://www.jstor.org/stable/42955888.

Elinder, M., Jordahl, H., & Poutvaara, P. (2015). Promises, policies and pocketbook voting. *European Economic Review*, *75*, 177–194. http://dx.doi.org/10.1016/j.euroecorev.2015.01.010.

Erikson, R. S., & Wlezien, C. (2012). *The timeline of presidential elections: how campaigns do (and do not) matter*. University of Chicago Press.

Erikson, R. S., & Wlezien, C. (2021). Forecasting the 2020 presidential election: Leading economic indicators, polls, and the vote. *PS: Political Science & Politics*, *54*(1), 55–58. http://dx.doi.org/10.1017/S1049096520001481.

Fair, R. C. (2009). Presidential and congressional vote-share equations. *American Journal of Political Science*, *53*(1), 55–72.

Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science*, *11*(1), 103–130. http://dx.doi.org/10.1561/100.00015031.

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. http://dx.doi.org/10.1016/j.ijforecast.2012.06.004.

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

Graefe, A. (2017). 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. (2019). Accuracy of german federal election forecasts, 2013 & 2017. *International Journal of Forecasting*, *35*(3), 868–877.

Graefe, A. (2021). Of issues and leaders: Forecasting the 2020 US presidential election. *PS: Political Science & Politics*, *54*(1), 70–72. http://dx.doi.org/10.1017/S1049096520001390.

Graefe, A., & Armstrong, J. S. (2012). Predicting elections from the most important issue: A test of the take-the-best heuristic. *Journal of Behavioral Decision Making*, *25*(1), 41–48. http://dx.doi.org/10.1002/bdm.710.

Graefe, A., & Armstrong, J. S. (2013). Forecasting elections from voters' perceptions of candidates' ability to handle issues. *Journal of Behavioral Decision Making*, *26*(3), 295–303. http://dx.doi.org/10.1002/bdm.1764.

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

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.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*(8), 1678–1685. http://dx.doi.org/10.1016/j.jbusres.2015.03.026.

Gruca, T. S., & Rietz, T. A. (2021). The 2020 (re)election according to the iowa electronic markets: Politics, pandemic, recession, and/or protests? *PS: Political Science & Politics*, *54*(1), 86–90. http://dx.doi.org/10.1017/S1049096520001419.

Haynes, S., & Stone, J. (2008). A disaggregate approach to economic models of voting in US presidential elections: forecasts of the 2008 election. *Economics Bulletin*, *4*(28), 1–11.

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. http://dx.doi.org/10.1111/j.1467-9280.2009.02271.x.

Hibbs, D. A. (2000). Bread and peace voting in U.S. presidential elections. *Public Choice*, *104*(1), 149–180. http://dx.doi.org/10.1023/A:1005292312412.

Holbrook, T. M. (2016). National conditions, trial-heat polls, and the 2016 election. *PS: Political Science & Politics*, *49*(4), 677–679. http://dx.doi.org/10.1017/S1049096516001347.

Hsieh, J. F.-s., Lacy, D., & Niou, E. M. S. (1998). Retrospective and prospective voting in a one-party-dominant democracy: Taiwan's 1996 presidential election. *Public Choice*, *97*(3), 383–399. http://dx.doi.org/10.1023/A:1005062527921.

Jerôme, B., & Jerôme-Speziari, V. (2016). State-level forecasts for the 2016 US presidential elections: Political economy model predicts hillary clinton victory. *PS: Political Science & Politics*, *49*(4), 680–686. http://dx.doi.org/10.1017/S1049096516001311.

Klarner, C. E. (2012). State-level forecasts of the 2012 federal and gubernatorial elections. *PS: Political Science & Politics*, *45*(4), 655–662. http://dx.doi.org/10.1017/S1049096512000960.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127.

Lewis-Beck, M. S., & Tien, C. (2012). Election forecasting for turbulent times. *PS: Political Science & Politics*, *45*(4), 625–629. http://dx.doi.org/10.1017/S1049096512000893.

Lewis-Beck, M. S., & Tien, C. (2021). The political economy model: A blue wave forecast for 2020. *PS: Political Science & Politics*, *54*(1), 59–62. http://dx.doi.org/10.1017/S1049096520001365.

Lichtman, A. J. (2008). The keys to the white house: An index forecast for 2008. *International Journal of Forecasting*, *24*(2), 301–309. http://dx.doi.org/10.1016/j.ijforecast.2008.02.004.

Lockerbie, B. (2021). Economic pessimism and political punishment in 2020. *PS: Political Science & Politics*, *54*(1), 67–69. http://dx.doi.org/10.1017/S1049096520001444.

Norpoth, H. (2014). The electoral cycle. *PS: Political Science & Politics*, *47*(2), 332–335. http://dx.doi.org/10.1017/S1049096514000146.

Norpoth, H. (2016). Primary model predicts trump victory. *PS: Political Science & Politics*, *49*(4), 655–658. http://dx.doi.org/10.1017/S1049096516001323.

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.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430. http://dx.doi.org/10.1002/for.928.

Thomson, M. E., Pollock, A. C., Önkal, D., & Gönül, M. S. (2019). Combining forecasts: Performance and coherence. *International Journal of Forecasting*, *35*(2), 474–484. http://dx.doi.org/10.1016/j.ijforecast.2018.10.006.