Assessing the 2016 U.S. Presidential Election Popular Vote Forecasts

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

J. Scott Armstrong
The Wharton School, University of Pennsylvania, Philadelphia, PA, and Ehrenberg-Bass Institute, University of South Australia, Adelaide, SA, Australia.
armstrong@wharton.upenn.edu

Randall J. Jones, Jr.
University of Central Oklahoma, USA
ranjones@uco.edu

Alfred G. Cuzán
University of West Florida, USA
acuzan@uwf.edu

Feb 7, 2017-R -Forthcoming in
The 2016 Presidential Election: The causes and consequences of an Electoral Earthquake,
Lexington Books, Lanham, MD

Abstract

The PollyVote uses evidence-based techniques for forecasting the popular vote in presidential elections. The forecasts are derived by averaging existing forecasts generated by six different forecasting methods. In 2016, the PollyVote correctly predicted that Hillary Clinton would win the popular vote. The 1.9 percentage-point error across the last 100 days before the election was lower than the average error for the six component forecasts from which it was calculated (2.3 percentage points). The gains in forecast accuracy from combining are best demonstrated by comparing the error of PollyVote forecasts with the average error of the component methods across the seven elections from 1992 to 2012. The average errors for last 100 days prior to the election were: public opinion polls (2.6 percentage points), econometric models (2.4), betting markets (1.8), and citizens' expectations (1.2); for expert opinions (1.6) and index models (1.8), data were only available since 2004 and 2008, respectively. The average error for PollyVote forecasts was 1.1, lower than the error for even the most accurate component method.

Introduction

The 2016 U.S. presidential election represented both a success and a failure for the forecasting community. Nearly every forecast predicted that Hillary Clinton would receive more votes than Donald Trump. Indeed, she received almost three million more votes than he did, the difference made up in only three metropolitan areas (Los Angeles, New York City, and the District of Columbia), for a 51.1-48.9% split in the two-party vote. But, of course, in the United States it is not the popular vote but the Electoral College, where states are represented somewhat less than in proportion to their population, which decides the issue. Historically, the two results have been at variance only a few times—and 2016 was one of those. Trump beat Clinton in several Midwestern states plus Pennsylvania by a combined total of about 100,000 votes, enough to win all of those states' electoral votes and carry the day in the Electoral College, 304-227. To the best of our knowledge, no forecast, including our own, anticipated this outcome, because of large polling errors in those very states.

In this chapter, we focus on forecasts of the popular vote, rather than the electoral vote. Ianalyze the accuracy of six different forecasting methods in predicting the popular vote in the 2016 U.S. presidential election, and then compare their performance to historical elections since 1992. These methods are based on people's vote intentions (collected by poll aggregators), people's expectations of who is going to win (evident in prediction markets, expert judgment, and citizen forecasts), and statistical models based on patterns estimated from historical elections (in econometric

models and index models). In addition, we review the performance of the PollyVote, a combined forecast based on these six different methods, and show why the PollyVote did not perform as well in 2016 as in previous years.

The PollyVote

The PollyVote research project was launched in 2004. The project's main goal was to apply evidence-based forecasting principles to election forecasting. That is, the purpose was to demonstrate that these principles – which were derived from forecasting research in different fields and generalized for forecasting in any field – could produce more accurate, and more useful, election forecasts. We view the PollyVote as useful in that its predictions begin early in elections years, in time to aid decision-making. Thus, we focus more on long term prediction, rather than election eve forecasts.

The PollyVote is a long-term project. The goal is to learn about the relative accuracy of different forecasting methods over time and in various settings. The PollyVote has now been applied to the four U.S. presidential elections from 2004 to 2016, as well as to the 2013 German federal election. In addition, the goal is to continuously track advances in forecasting research and apply them to election forecasting. This has led to the development of index models, which are particularly well suited to aiding decisions by campaign strategists, and to validating previous work on citizen forecasts, an old method that has been widely overlooked despite its accuracy (Graefe 2014).

Combining forecasts

At the core of the PollyVote lies the principle of combining forecasts, which has a long history in forecasting research (Armstrong 2001). Combining evidence-based forecasts – forecasts from methods that have been validated for the situation – has obvious advantages.

First, any one method or model is limited in the amount of information that it can include. Because the resulting forecast does not incorporate all relevant information, it is subject to bias. Combining forecasts from different methods that use different information helps to overcome this limitation.

Second, forecasts from different methods and data tend to be uncorrelated and often bracket the true value, the one that is being predicted. In this situation both systematic and random errors of individual forecasts tend cancel out in the aggregate, which reduces error.

Third, the accuracy of different methods usually varies across time, and methods that have worked well in the past often do not perform as well in the future. Combining forecasts thus prevents forecasters from picking a poor forecast.

Mathematically, the approach *guarantees* that the combined forecast will at least be as accurate as the typical component forecast.[1] Under ideal conditions, and when applied to many

[1] The error of the typical component is the average error of the individual components. That is, it represents the error that one would get by randomly picking one of the available component forecasts.

forecasting problems, a combined forecast often outperforms even its most accurate component (Graefe et al. 2014b).

Conditions for combining forecasts

While combining is useful whenever more than one forecast for the same outcome is available, the approach is particularly valuable if many forecasts from evidence-based methods are available and if the forecasts draw upon different methods and data (Armstrong 2001). These conditions apply to election forecasting (Graefe et al. 2014b). First, there are many evidence-based methods for predicting election outcomes, including the six that comprise the PollyVote, noted previously (polls, prediction markets, expert judgment, citizen forecasts, econometric models, and index models). Second, these methods rely on different data.

Although the reasoning that underlies these two conditions may be self-evident, the value of the combined forecast is less clear, primarily because many people wrongly believe that combining yields only average performance (Larrick and Soll 2006), which is the worst possible outcome for a combined forecast. People subject to that misperception often try to identify the best component forecast, but then pick a poor forecast that is less accurate than the combined one (Soll and Larrick 2009).

How to best combine forecasts

A widespread concern when combining forecasts is how best to weight the components, and scholars have tested various weighting methods. However, a large literature suggests that the simple average, assigning equal weights to the components, often provides more accurate forecasts than complex approaches, such as assigning "optimal" weights to the components based on their past performance (Graefe et al. 2015, Graefe 2015c).

One reason for the accuracy of equal weights is that the relative accuracy of component forecasts varies over time. For example, when analyzing the predictive performance of six econometric models across the ten U.S. presidential elections from 1976 to 2012, one study found a negative correlation between a model's past and future performance (Graefe et al. 2015). In other words, models that were among the most accurate in a given election tended to be among the least accurate in the succeeding election. Obviously, in this circumstance weighting the forecasts based on past performance is unlikely to produce accurate combined forecasts.

More important than the decision of how to weight the components is the timing of that decision. In particular, forecasters must not make the decision as to the method of combining components at the time they are making the forecasts. This is because they may then weight the components in a way that suits their biases. To prevent that, the combining procedure should be specified before generating the forecasts and should not be adjusted afterwards.

The combined PollyVote forecast

The PollyVote combines numerous forecasts from several different forecasting methods, each of which relies on different data. The optimal conditions for combining, identified by Armstrong

(2001), are thus met. In 2016, the PollyVote averaged forecasts within and across six different component methods, each of which has been shown in research findings to be a valid method for forecasting election outcomes.

While the number of component forecasts has increased since the PollyVote's first launch in 2004, the two-step approach for combining forecasts has remained unchanged. We first average forecasts *within* each component method and then average the resulting forecasts *across* the component methods. In other words, weighing them equally, we average the forecasts within each method; then, again using equal weights, we average the within-method averages across the different methods. This is the same approach that the PollyVote has successfully used to forecast U.S. presidential elections in 2004 (Cuzán, Armstrong, and Jones 2005), 2008 (Graefe et al. 2009), and 2012 (Graefe et al. 2014a), as well as the 2013 German federal election (Graefe 2015b).

The rationale behind choosing this two-step procedure is to equalize the impact of each component method, regardless whether a component includes many forecasts or only a few. For example, while there is only one prediction market that predicts the national popular vote in U.S. presidential elections, there are forecasts from numerous econometric models. In this situation, a simple average of all available forecasts would over-represent models and under-represent prediction markets, which we expect would reduce the accuracy of the combined forecast. Thus, the one prediction market is weighted equally with the average forecast of all econometric models.

Past performance

The 2004 PollyVote was introduced in March of that year. The original specification combined forecasts from four methods: polls, prediction markets, expert judgment, and econometric models. The PollyVote predicted a popular vote victory for President George W. Bush over the eight months that it was producing forecasts. The final forecast, published on the morning of the election, predicted that the President would receive 51.5% of the two-party popular vote, an error of 0.3 percentage points (Cuzán, Armstrong, and Jones 2005).

Using the same specification as in 2004, the 2008 PollyVote commenced in August 2007. It forecast a popular vote victory for Barack Obama over the 14 months that it was making daily forecasts. On Election Eve the PollyVote predicted that Obama would receive 53.0% of the popular two-party vote, an error of 0.7 percentage points (Graefe et al. 2009).

The 2012 PollyVote was launched in January 2011 and forecast a popular vote victory for President Obama over the 22 months that it was making daily forecasts. On Election Eve, it predicted that Obama would receive 51.0% of the popular two-party vote, an error of 0.9 percentage points. This was also the first year that index models were added as a separate component (Graefe et al. 2014a).

An ex post analysis tested how the PollyVote would have performed since 1992 by adding three more elections to the data set, 1992, 1996, and 2000. Across the last 100 days prior to Election Day, on average the PollyVote provided more accurate popular vote forecasts than each of the

component methods. Error reductions were large. For example, compared to single polls, the PollyVote reduced forecast error by 59% (Graefe et al. 2014b).

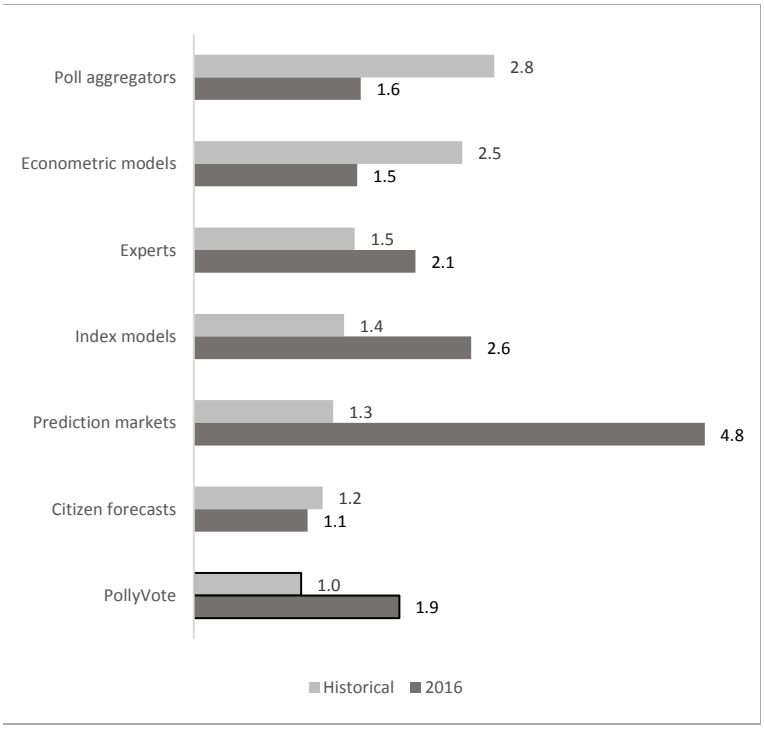
In addition, the PollyVote was used to predict the vote shares of six parties in the 2013 German Federal election by combining forecasts from polls, prediction markets, econometric models, and expert judgment. On average, across the two months prior to the election, which is the maximum time frame for 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 2015b).

2016 forecast accuracy

Since its first appearance in January of 2016, the combined PollyVote consistently – and correctly – predicted Hillary Clinton to win the popular vote. However, with a mean absolute error (MAE) of 1.9 percentage points across the last 100 days before the election, the forecast error was almost twice as large as the corresponding average error across the six previous elections from 1992 to 2012, which was only 1.0 percentage point (cf. Figure 1).[2] Moreover, on average in previous elections the PollyVote was more accurate than each of its components. This was not the case in 2016. In the remainder of this chapter, we assess each component method's performance in 2016 and discuss why the PollyVote did not perform as well in this election.

Figure 1. Forecast error by method

(Mean absolute error, historical (1992-2012) vs. 2016, across last 100 days before the election)



[2] The MAE across the last 100 days prior to the election is determined in the following manner: First, we calculate a method's error for each of the 100 days as the absolute difference between the predicted and actual election outcome. Second, we average the daily errors across the 100-day period.

Methods for forecasting elections

The following section reviews the accuracy of the six different component methods included in the PollyVote for predicting the 2016 popular vote, and compares the resulting errors to the methods' historical performance.

Polls

Trial heat polls are the most prevalent election predictors and are and highly visible in news media coverage. With this method interviewers ask respondents a variation of this question*: "If the election for President were held today, for whom would you vote: Donald Trump, the Republican, or Hillary Clinton, the Democrat?"* Note that respondents are asked to state their candidate choice if the election were held today. Thus polls do not provide predictions; they provide snapshots of public opinion at a certain point in time.

However, this is not how the media commonly treat polls. Polling results are routinely interpreted as forecasts of candidate performance on Election Day. This interpretation of polls can result in poor predictions, especially when the election is still far into the future, because polls tend to vary widely over the course of the campaign (Gelman and King 1993). Also, there is often high variance in the results of polls conducted at about the same time by different survey organizations. This variation can be caused by sampling problems, nonresponses, faulty processing and other sources of bias (Erikson and Wlezien 2012).

It is apparent, therefore, that one should not rely on the results of a single poll. Rather, one should combine polls that were conducted near the same time, since the errors associated with individual polls tend to cancel out in the aggregate (Graefe et al. 2014b). Systematic error, however, may persist due to nonresponse, for example, which often points poll results in the same direction.

The public's increasing awareness of the variation in poll results and the value of combining have had a positive impact on the way people consume polls. Online poll aggregators, such as realclearpolitics.com, pollster.com, and fivethirtyeight.com have become increasingly popular.

The 2016 PollyVote relied on several poll aggregators, each of which used different methods to collect and combine individual polls. Aggregators commonly differ in their policies as to which polls to include in their averages, how to weight them, and how transparent to be about their methodology. To calculate its combined poll component, the PollyVote averaged daily forecasts of the chosen poll aggregators. Figure 2 shows the PollyVote's combined polls forecast of Hillary Clinton's two-party vote. The horizontal axis depicts Clinton's actual two-party vote (51.1%). Except for a brief three-day period in mid-September, combined poll forecasts never fell below Clinton's actual vote, and at times exceeded it by almost three percentage points.

Figure 3. Combined polls forecast 2016

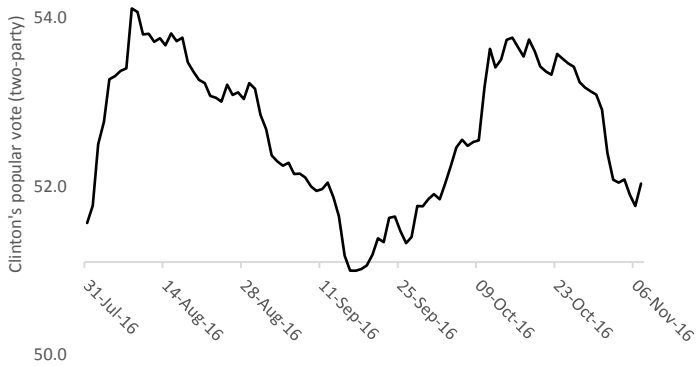
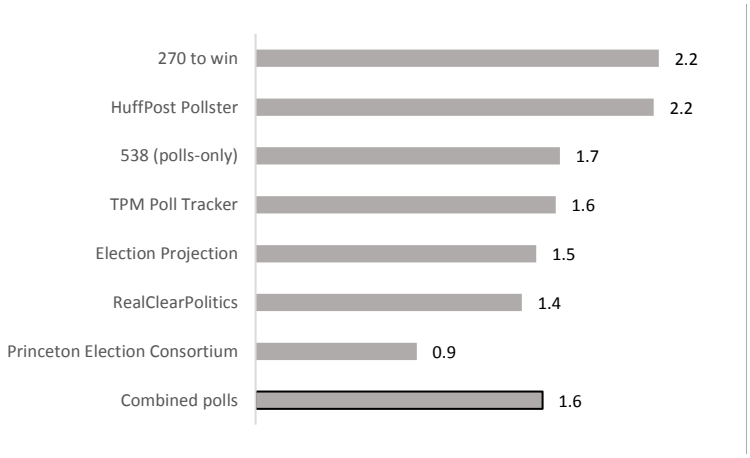


Figure 3 shows the mean absolute error of each individual poll aggregator, as well as the PollyVote's combined polls component, across the last 100 days before the election.[3] The error of the different polling aggregators ranged from 0.9 percentage points for Sam Wang's Princeton Election Consortium to 2.2 percentage points for both the HuffPost Pollster and 270 To Win (cf. Figure 2). With an error of 1.6 percentage points, the PollyVote's combined poll component performed as well as the typical poll aggregator.

Figure 3. Forecast error of poll aggregators

(Mean absolute error, across last 100 days before the 2016 election)



As shown in Figure 1, national polls were considerably more accurate than in previous elections from 1992 to 2012, when the corresponding error was on average 2.8 percentage points.

[3] The PollyVote's combined polls component also included data from the NYT poll average and YouGov. However, Figure 2 does not show the errors from these aggregators since we did not have data across the complete 100-day period.

Expectations-based methods

Three of the PollyVote indicators reflect individuals' efforts to project the election winner and thereby form expectations of the election outcome. Panels of experts who are engaged to forecast the election winner form such expectations. Bettors who wager in election prediction markets, and even the general public, also develop expectations as to which candidate will win. For the analyst, identifying these expectations provides a means of predicting the election outcome.

Expert judgment

Asking experts to predict what is going to happen is one of the oldest forecasting methods. With regard to elections, experts in that subject can be expected to have broad knowledge of the campaign and electoral process, as well as expertise in interpreting measures of the status of a campaign, such as reflected in polls. Because their opinion is better informed than that of the public, one might expect their judgment to be more accurate than polls. Some evidence suggests that this is the case. Jones and Cuzán (2013) found that experts provided more accurate forecasts than polls early in the election season, when the election was still at least nine months in the future.

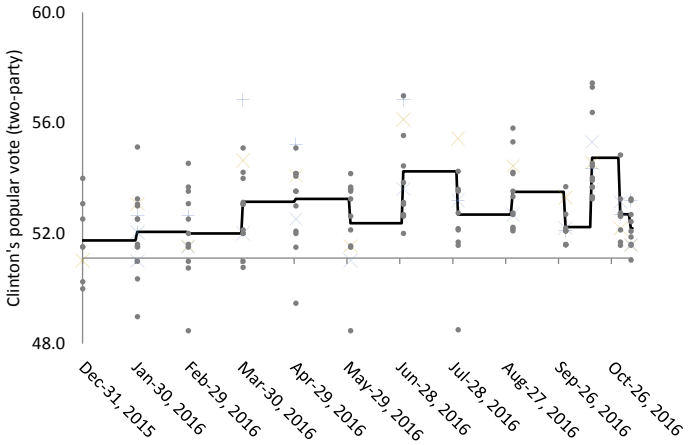
The PollyVote includes the judgment of prominent academics (and in 2004 some practitioners, as well) who are knowledgeable of American politics. In 2016, a panel of 15 political scientists[4] was polled 13 times between late December 2015 and Election Day. The mean forecast was incorporated into the PollyVote.

Figure 4 shows the individual (grey dots) and combined (black line) expert forecasts of Hillary Clinton's two-party vote. The horizontal axis depicts Clinton's actual two-party vote. Except for the very first survey conducted in December 2015, Clinton's predicted two-party vote in the combined expert forecast never fell below 52.0%, and was always above her final vote share of 51.1%. In other words, the experts consistently over-predicted Clinton's vote share.

In our period of analysis, the last 100 days before the election, we conducted six surveys with a total of 77 individual expert forecasts, 75 of which overestimated the vote share Clinton would eventually receive. Across the last 100 days, the mean absolute error of the combined expert forecast was 2.1 percentage points (cf. Figure 1). This error is 40% higher than the experts' corresponding error across the three elections from 2004 to 2012, which was 1.5 percentage points.

Figure 4. Individual and combined expert forecasts 2016

[4] Thanks to Randall Adkins (University of Nebraska, Omaha), Lonna Rae Atkeson (University of New Mexico), Scott Blinder (University of Massachusetts, Amherst), John Coleman (University of Minnesota), George Edwards (Texas A&M University), John Geer (Vanderbilt University), Sandy Maisel (Colby College), Michael Martinez (University of Florida), Thomas Patterson (Harvard University), Gerald Pomper (Rutgers University), David Redlawsk (Rutgers University), Larry Sabato (University of Virginia), Michael Tesler (University of California, Irvine), Charles Walcott (Virginia Tech), and one expert who preferred to remain anonymous. Originally, the panel consisted of 17 experts, two of which dropped out after the second survey, conducted in January 2016.



Prediction markets

Prediction markets are another expression of expectations as to who will win an election. Participants in prediction markets reveal their opinion by betting money on the election outcome. The price at which trades are made provides a forecast of a given candidate's vote share. Depending on the accuracy of their individual predictions, participants can either win or lose money, and thus have an incentive to be right. Hence, savvy bettors know to participate only if they believe they have information that improves the current market forecast. Generally, anyone may place bets in the markets, so there is no random sampling in choosing participants.

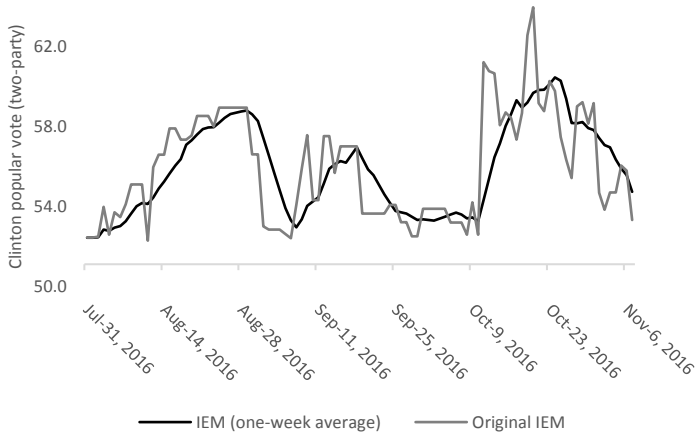
Graefe (2017) reviewed prediction market accuracy of vote-share forecasts for elections in different countries. He found that prediction markets tend to outperform forecasts made by experts, as well as forecasts based on quantitative models and trial-heat polls, although compared to citizen forecasts the evidence was mixed.

Most available markets provide probability forecasts for the candidates' likelihood to win and are thus not suitable for the PollyVote, which requires forecasts of the national popular vote shares. We know of only one prediction market that provides such information, the Iowa Electronic Market (IEM) at the University of Iowa. However, this market is of limited value due to a lack of efficiency. That is, the IEM has relatively low volume, and participants are not allowed to invest more than $500. The PollyVote uses the IEM's daily market prices, but calculates one-week rolling averages to limit short-term fluctuations. Across the six elections from 1992 to 2012, this procedure reduced the error of the original IEM forecasts by 10% on average (Graefe et al. 2014b).

Figure 5 shows Clinton's two-party popular vote forecasts from the original IEM and the PollyVote's one-week average across the last 100 days prior to the 2016 election. The horizontal axis depicts the actual election results. The figure shows that the IEM consistently, and at times dramatically, overestimated Clinton's vote share. On average, the one-week average of the IEM

missed the final election outcome by 4.8 percentage points[5], which makes it by far the least accurate component method in 2016.

Figure 5. Prediction market forecasts 2016



The weak performance of the IEM in 2016 is in stark contrast to the method's historically high accuracy. Across the six elections from 1992 to 2012, the IEM was the second most accurate among the PollyVote's components, with a MAE of merely 1.3 percentage points (cf. Figure 1).

We can only speculate as to the reasons why the IEM failed dramatically in 2016. One explanation could be systematic bias among the market participants. Prior research shows that IEM participants tend to be well educated, to belong to middle and upper income groups, and to be more politically interested and engaged (Forsythe et al. 1992). In other words, IEM participants are likely upscale in socioeconomic status, which may have resulted in anti-Trump preferences within this group, opposing the brash, coarse candidate of working-class white males.

Citizen forecasts

Vote expectation surveys—or citizen forecasts—are the newest addition to the PollyVote. Vote expectation surveys ask respondents who they expect to win the election, rather than asking them for whom they themselves intend to vote (Hayes 1936). A typical question might be: *"Who do you think will win the U.S. presidential election, Donald Trump or Hillary Clinton?"* The aggregate responses are then used to predict the election winner.

Though often overlooked, these citizen forecasts are highly accurate predictors of election outcomes (Graefe 2014). In 89% of 217 surveys administered between 1932 and 2012, a majority of respondents correctly predicted the winner. Regressing the incumbent share of the two-party vote on the percent of respondents who expect the incumbent party ticket to win accounts for two-thirds of the variance. Moreover, in the last 100 days of the previous seven presidential elections, vote expectations provided more accurate forecasts than vote intention polls, prediction markets, econometric models, and expert judgment. Compared to a typical poll, for example, vote expectations reduced the forecast error by about 50% on average. Furthermore, an ex post analysis for the elections from 1992 to 2012

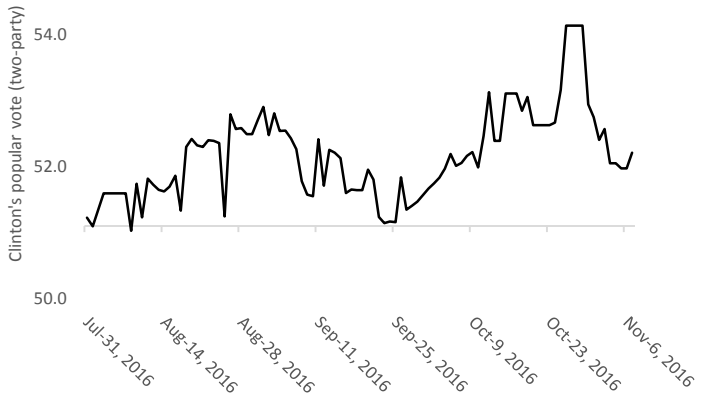
[5] Contrary to previous elections, calculating one-week averages only marginally improved the accuracy of the IEM in 2016 (error reduction: 1%).

found that adding citizen forecasts to the PollyVote would have reduced forecast error by another 7% (Graefe 2015a).

Across the last 100 days prior to the 2016 election, we collected 39 surveys that asked people who they think will win the election, plus daily data starting on August 8 from the Reuters tracking poll. We translated the results of each individual vote expectation survey into a two-party vote share prediction using the vote equation estimated by Graefe (2014). We then averaged the forecasts of the most recent survey from all other established sources and the most recent Reuters data to calculate the PollyVote's combined citizen component forecast.

Figure 6 shows the PollyVote's daily citizen forecasts of Clinton's two-party popular vote across the last 100 days before the election. As in previous figures, the horizontal line depicts the actual election result. The citizen forecast constantly overestimated Clinton's vote share, particularly in the month of October, but forecast errors were low. In fact, citizen forecasts were the most accurate method for predicting the 2016 popular vote. Across the final 100 days before the election, citizen forecasts on average missed by only 1.1 percentage points. The method thus once again demonstrated its high level of accuracy, as in previous elections. As shown in Figure 1, the average error of citizen Please expain what was averaged and how (i.e., how many times were the forecasrts conducted and with what sample size each time? (JSA) forecasts across the last 100 days for the six elections from 1992 to 2012 was only 1.2 percentage points.

Figure 6. Combined citizen forecasts 2016



Models

In addition to other indicators, the PollyVote combined forecasts from two types of models: indexes and econometric models. The two are quite dissimilar, however, in their underlying theory and the data upon which they rely.

Econometric models

For the past several presidential election cycles at least a dozen political scientists and economists have computed regression equations to forecast the election results. Many of the models

use economic growth data through the second quarter of the election year, the first official estimate of which becomes available in late July. Most forecasts from those models are made shortly after that.

However, the predictions of some models are available well before then, even years ahead of the election (Norpoth 2014), while at least one is delayed until the first polls after Labor Day are released (Campbell 2016). Also, while most models provide a single prediction, others, such as FiveThirtyEight, are updated almost daily, as when new polls become available.

Most of these models are based on the theory of retrospective voting. This concept assumes that, in casting the ballots, voters assess the performance of the incumbent party, particularly in handling the economy. A good performance is rewarded with another term in office. In addition, many models include some measure of the length of time that the incumbent party has been in office, which recognizes the public's periodic desire for change. Some models also include an indicator of the president's popularity.

In 2016 these models on average predicted a very close race. Their mean forecast across the last 100 days pointed to a virtual tie, predicting that Clinton would receive 49.6% of the popular two-vote. That said, there was a wide spread in the 10 models' individual forecasts, which differed by as much as 10-points, ranging from 44.0% (Fair) to 53.9% (Hibbs) of the two-party vote.

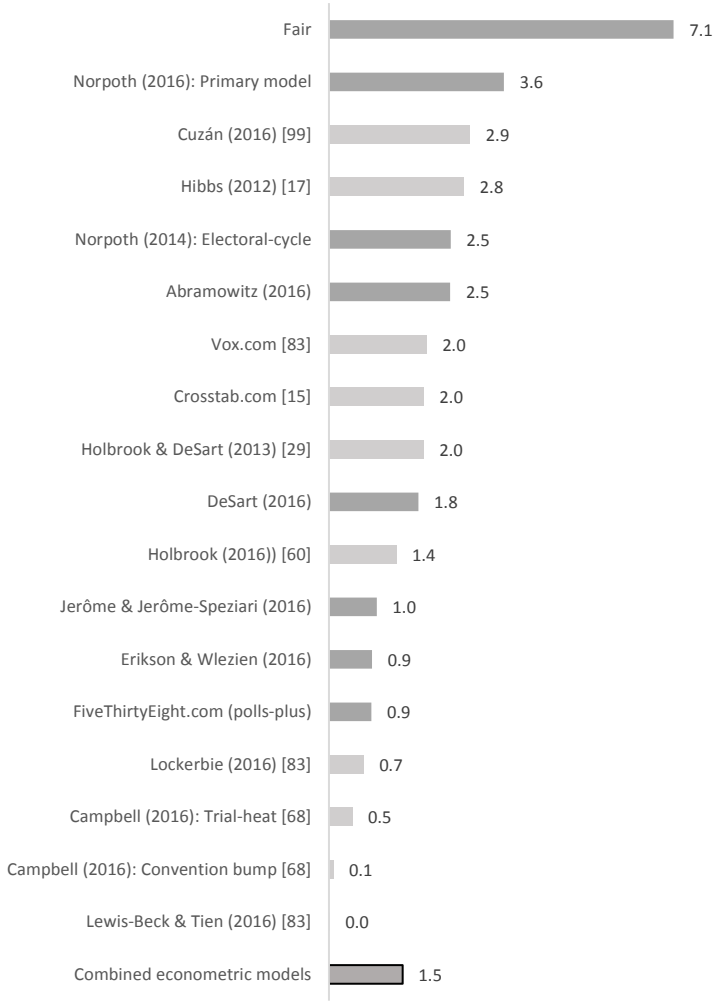
Figure 7 shows the MAE for each individual model across the last 100 days before the election.[6] The light grey bars represent models for which forecasts have not been available for the complete 100-day period. The numbers in parentheses shows the number of days before the election when the first forecast from that model became available. The model by Yale economist Ray Fair incurred the largest error with 7.1 percentage points, while the Lewis-Beck and Tien model predicted the outcome perfectly.

Across the last 100 days before the election the forecast of the combined econometric models missed the final election outcome by 1.5 percentage points, which makes it more accurate than 10 of the 18 individual models. Also, the econometric model component was the second most accurate component method in forecasting the 2016 election, after citizen forecasts.

[6] In addition to the three forecasts published at fivethirtyeight.com, vox.com, and crosstab.com, please refer to the respective publications for details about each model (Erikson and Wlezien 2016, Campbell 2016, Fair 2009, Norpoth 2014, 2016, Hibbs 2012, Abramowitz 2016, Cuzán 2016, Holbrook 2016, Jerôme and Jerôme-Speziari 2016, Lockerbie 2016, Lewis-Beck and Tien 2016, DeSart 2016, Holbrook and DeSart 2013).

Figure 7. Forecast error of econometric models

(Mean absolute error, across last 100 days before the 2016 election)



Index models

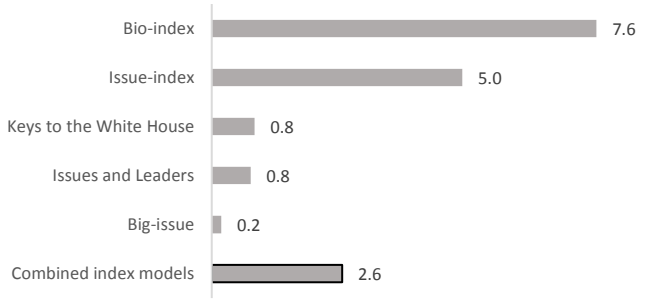
In contrast to econometric models, most index models are based on the concept of prospective voting. These models assume that voters assess the personal traits of the candidates and their positions on important issues, when deciding for whom to vote. Indexes are typically constructed from ratings of specific characteristics of candidates or events. Ratings can be made by experts or members of the public (as in survey data) and can cover factors such as the candidates' biographic information, leadership skills, or issue-handling competences, as well as exogenous effects, such as economic performance or the presence of a third party. Point forecasts of an election are provided by inserting current data into an equation specified by regressing the vote on the respective index scores.

As shown in Figure 8, the five available index models overestimated Clinton's support by an average of 2.6 percentage points, primarily due to the large error of two models, the bio-index (Armstrong and Graefe 2011) and the issue-index (Graefe and Armstrong 2013). In comparison, the

three remaining models-- including the big-issue model (Graefe and Armstrong 2012), the Issues and Leaders model (Graefe 2013), and the Keys to the White House (Lichtman 2008) -- were quite close to the final election outcome.

Figure 8. Forecast error of index models

(Mean absolute error, across last 100 days before the 2016 election)



Discussion

Prior research shows that the relative accuracy of different forecasting methods varies significantly from one election to the next. This was true again in 2016. Prediction markets, which have been among the most accurate methods historically, were off dramatically, while econometric models, which historically have had rather high error, were more accurate in 2016. In fact, based on data from Figure 1, a negative correlation exists between the methods' accuracy in previous elections and their performance in 2016 (r=-.4).

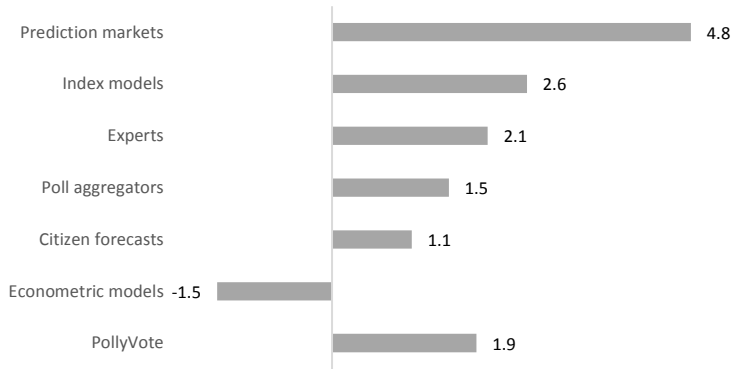
In other words, it is extremely difficult to foresee which method will be the most (or least) accurate in a given election. This is, of course, one of the major reasons why combining forecasts is such a useful strategy. The combined forecast protects one from making large mistakes that can occur when relying on a single poor forecast.

Combining works best when the errors of individual forecasts are uncorrelated. Then, the true value lies near the midpoint of the range of the various component forecasts, a situation commonly referred to as bracketing (Graefe et al. 2014b). Under ideal conditions, bracketing can result in a situation where the combined forecast outperforms the most accurate component method when forecasting a single election. In the case of the PollyVote, this happened in 2004 (Cuzán, Armstrong, and Jones 2005) and 2012 (Graefe et al. 2014a).

In 2016, little bracketing occurred. As shown in Figure 9, five of the six components consistently predicted Clinton's share of the vote to be higher than it was. Only one component, the econometric models, underestimated the Clinton vote. As a result, the PollyVote did not perform as well as in previous elections and was only slightly more accurate than the typical forecast. The PollyVote outperformed expert judgment, index models, and prediction markets, but performed worse than econometric models, citizen forecasts, and combined polls.

Figure 9. Direction of errors by component method

(average error across last 100 days before the election; -: under-predicted Clinton; +: over-predicted Clinton)



It is noteworthy that the polls and all three methods that rely on expectations (prediction markets, expert judgment, and citizen forecasts) over-predicted Clinton's vote share. Experts – including self-selected experts in prediction markets – apparently thought that the polls would underestimate Clinton, and tended to assign even higher numbers for her anticipated vote share. In retrospect, perhaps the experts' forecasts were influenced by factors such as Clinton' consistent lead in the polls, her large post-convention bounce, the consistently bad coverage that Trump received in the elite press read by academics, and Trump's unconventional campaign.

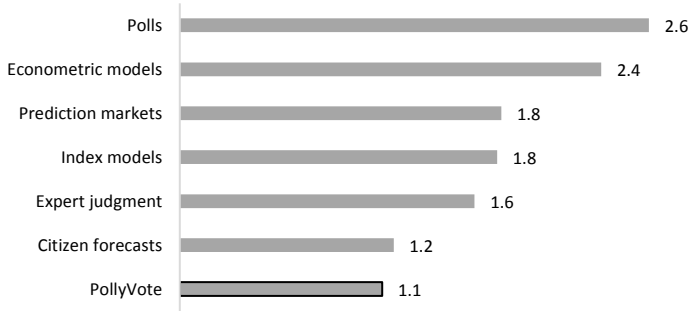
However, as shown by the econometric models component, which always predicted a very tight race, there was also information that pointed in the other direction. Of the nine models that did not include trial-heat polls, the average forecast was that Trump would win 50.8% of the popular vote.[7] Of course, this forecast was also wrong because Clinton won the popular vote. But it does show that information existed which could have alerted the close observer of econometric models that Clinton's anticipated vote may have been be over- (rather than under-) estimated in the polls.

When most component forecasts err in the same direction, as in 2016, the combined forecast will perform only slightly better than the typical forecast. But the principle of combining does not claim that the combined forecast will always outperform its most accurate component. Yet, over time, as the component methods' relative accuracy varies, the combined forecast likely will surpass them. This is shown in Figure 10, which depicts the mean absolute error for the last 100 days before each election from 1992 to 2016. By including 2016, the data incorporate the most recent observations for each component method and the combined PollyVote forecast. The PollyVote's MAE of 1.1 percentage points is lower than the corresponding error of any other method.

Figure 10. Forecast error by method

(Mean absolute error, 1992-2016, across last 100 days before the election)

[7] The nine models are those by Abramowitz (2016), Cuzán (2016), Fair (2009), Hibbs (2012), Jerôme and Jerôme-Speziari (2016), Lewis-Beck and Tien (2016), Lockerbie (2016), Norpoth (2014), Norpoth (2016).



Conclusion

At the PollyVote, we are always reviewing types of forecasts to include, methods of combining the forecasts, and especially means of measuring their uncertainty. In addition, we are constantly looking for new research evidence that can improve the accuracy of our forecasts.

That said, forecasts for one election should not cause us to doubt fundamental principles from nearly half a century of forecasting research. We know that the combined forecast will always be at least as accurate as the typical component forecast in any single event. As a consequence, we also know that the principle of combining forecasts prevents the forecaster from making large errors. We further know that the performance of individual forecasts varies widely over time and in different settings. Combined forecasts, therefore, will be among the most accurate forecasts available. In the long run, there is no better way to forecasting than by combining different methods that use different methods with different information.

References

Abramowitz, Alan I. 2016. "Will Time for Change Mean Time for Trump?" *PS: Political Science & Politics* 49 (4):659-660. doi: 10.1017/S1049096516001268.

Armstrong, J. Scott. 2001. "Combining forecasts." In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, edited by J. Scott Armstrong, 417-439. New York: Springer.

Armstrong, J. Scott, and Andreas Graefe. 2011. "Predicting elections from biographical information about candidates: A test of the index method." *Journal of Business Research* 64 (7):699-706. doi: 10.1016/j.jbusres.2010.08.005.

Campbell, James 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. doi: 10.1017/S104909651600127X.

Cuzán, Alfred G. 2016. "Fiscal model forecast for the 2016 presidential election." *SSRN Working Paper*:https://ssrn.com/abstract=2821878.

Cuzán, Alfred G., J. Scott Armstrong, and Randall J. Jones, Jr. 2005. "How we Computed the PollyVote." *Foresight: The International Journal of Applied Forecasting* 1 (1):51-52.

DeSart, Jay A. 2016. "A Long-Range, State-Level Presidential Election Forecast Model." 2016 Annual Meeting of the American Political Science Association, Philadelphia, PA, September 1-4.

Erikson, Robert S., and Christopher Wlezien. 2012. *The Timeline of Presidential Elections: How Campaigns Do (And Do Not) Matter*. Chicago: University of Chicago Press.

Erikson, Robert S., and Christopher Wlezien. 2016. "Forecasting the Presidential Vote with Leading Economic Indicators and the Polls." *PS: Political Science & Politics* 49 (4):669-672. doi: 10.1017/S1049096516001293.
Fair, Ray C. 2009. "Presidential and congressional vote-share equations." *American Journal of Political Science* 53 (1):55-72.
Forsythe, Robert, Forrest Nelson, George R. Neumann, and Jack Wright. 1992. "Anatomy of an experimental political stock market." *The American Economic Review* 82 (5):1142-1161.
Gelman, Andrew, and Gary King. 1993. "Why are American presidential election campaign polls so variable when votes are so predictable?" *British Journal of Political Science* 23 (4):409-451.
Graefe, Andreas. 2013. "Issue and leader voting in U.S. presidential elections." *Electoral Studies* 32 (4):644-657. doi: http://dx.doi.org/10.1016/j.electstud.2013.04.003.
Graefe, Andreas. 2014. "Accuracy of vote expectation surveys in forecasting elections." *Public Opinion Quarterly* 78 (S1):204-232. doi: 10.1093/poq/nfu008.
Graefe, Andreas. 2015a. "Accuracy gains of adding vote expectation surveys to a combined forecast of US presidential election outcomes." *Research & Politics* 2 (1):1-5. doi: 10.1177/2053168015570416.
Graefe, Andreas. 2015b. "German election forecasting: Comparing and combining methods for 2013." *German Politics* 24 (2):195-204. doi: 10.1080/09644008.2015.1024240.
Graefe, Andreas. 2015c. "Improving forecasts using equally weighted predictors." *Journal of Business Research* 68 (8):1792–1799. doi: 10.1016/j.jbusres.2015.03.038.
Graefe, Andreas. 2017. "Political Markets." In *SAGE Handbook of Electoral Behavior,*, edited by Kai Arzheimer, Jocelyn Evans and Michael S. Lewis-Beck, in press. SAGE.
Graefe, Andreas, and J. Scott Armstrong. 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.
Graefe, Andreas, and J. Scott Armstrong. 2013. "Forecasting elections from voters' perceptions of candidates' ability to handle issues." *Journal of Behavioral Decision Making* 26 (3):295-303.
Graefe, Andreas, J. Scott Armstrong, Randall J. Jones, Jr., and Alfred G. Cuzán. 2009. "Combined Forecasts of the 2008 Election: The PollyVote." *Foresight: The International Journal of Applied Forecasting* 2009 (12):41-42.
Graefe, Andreas, J. Scott Armstrong, Randall J. Jones, Jr., and Alfred G. Cuzán. 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, Andreas, J. Scott Armstrong, Randall J. Jones, Jr., and Alfred G. Cuzán. 2014b. "Combining forecasts: An application to elections." *International Journal of Forecasting* 30 (1):43-54.
Graefe, Andreas, Helmut Küchenhoff, Veronika Stierle, and Bernhard Riedl. 2015. "Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems." *International Journal of Forecasting* 31 (3):943-951.
Hayes, Samuel P. Jr. 1936. "The predictive ability of voters." *Journal of Social Psychology* 7 (2):183-191.
Hibbs, Douglas A. 2012. "Obama's reelection prospects under “Bread and Peace” voting in the 2012 US Presidential Election." *PS: Political Science & Politics* 45 (4):635-639.
Holbrook, Thomas M. 2016. "National Conditions, Trial-heat Polls, and the 2016 Election." *PS: Political Science & Politics* 49 (4):677-679. doi: 10.1017/S1049096516001347.
Holbrook, Thomas M., and Jay A. DeSart. 2013. "Forecasting Hurricanes and Other Things: The DeSart and Holbrook Forecasting Model and the 2012 Presidential Election." 2013 Annual Meeting of the Western Political Science Association, Hollywood, CA, March 28-30.
Jerôme, Bruno, and Véronique Jerôme-Speziari. 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. doi: 10.1017/S1049096516001311.
Jones, Randall J. Jr., and Alfred G. Cuzán. 2013. "Expert Judgment in Forecasting American Presidential Elections: A Preliminary Evaluation." Annual Meeting of the American Political Science Association (APSA), Chicago.
Larrick, Richard P., and Jack B. Soll. 2006. "Intuitions about combining opinions: Misappreciation of the averaging principle." *Management Science* 52 (1):111-127.

Lewis-Beck, Michael S., and Charles Tien. 2016. "The Political Economy Model: 2016 US Election Forecasts." *PS: Political Science & Politics* 49 (4):661-663. doi: 10.1017/S1049096516001335.

Lichtman, Allan J. 2008. "The keys to the White House: An index forecast for 2008." *International Journal of Forecasting* 24 (2):301-309.

Lockerbie, Brad. 2016. "Economic Pessimism and Political Punishment." *PS: Political Science & Politics* 49 (4):673-676. doi: 10.1017/S104909651600130X.

Norpoth, Helmut. 2014. "The Electoral Cycle." *PS: Political Science & Politics* 47 (2):332-335. doi: 10.1017/S1049096514000146.

Norpoth, Helmut. 2016. "Primary Model Predicts Trump Victory." *PS: Political Science & Politics* 49 (4):655-658. doi: 10.1017/S1049096516001323.

Soll, Jack B., and Richard P. Larrick. 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.