Forecasting the 2018 Midterm Election using National Polls and District Information

Joseph Bafumi, Dartmouth College

Robert S. Erikson, Columbia University

Christopher Wlezien, University of Texas at Austin

Abstract

We forecast party control of the US House of Representatives after the 2018 midterm election. First, we model the expected national vote relying on available generic Congressional polls and the party of the president. Second, we model the district vote based primarily on results from 2016 and the national swing. To produce our forecast, we incorporate the uncertainty in the predictions using 3,000 simulations of the national vote and district outcomes. Based on our analysis, the Democrats are projected to win a solid plurality of the national vote, above 53% of the two-party share, and gain control of the House with a narrow 7-seat majority. Our simulations yield considerable variation, however, with the Republicans winning the majority of seats 46% of the time but with the distinct possibility of a big Democratic wave.

In this paper, we present a forecast of the 2018 midterm US House election based on information available four months in advance. The approach builds on our forecasts of the 2006 (Bafumi, et al 2008), 2010 (Bafumi, et al 2010), and the 2014 (Bafumi, et al 2014) elections.[1] We incorporate information about the national forces at work in the election, which are evident early in the election year from generic congressional polls plus the party of the president. We also incorporate information about the 435 districts based mostly on the vote in 2016. To forecast the 2018 election, we simulate the national vote and district outcomes, details about which we provide below. The approach can be updated on a weekly or even daily basis based on changes in generic ballot polls.

We model the expected national vote and the manner in which the vote margin translates into seats. Each has an element of uncertainty that we factor into our forecast.

The modeling of district voting carries particular risks beyond the model at hand. The model makes certain assumptions about the parameters of the district vote from the previous election (here, 2016) carrying over to the next (2018). This never happens exactly.

More importantly, the model assumes that district-level party effort carries forward from one election to the next, in that the parameters of the vote model are not conditional on shifting percepts of district competitiveness. In other words, it does not take into account extra effort by the Democratic Party in what are thought to be newly competitive districts. This limitation suggests a likelihood of underestimating any coming Democratic wave. Thus, the modeling here might be best considered a lower bound to the likely fortunes of the Democratic Party in the 2018 House elections.

With that note of caution, here is our forecast. Based on information gathered from June, we expect a very close race for the House. In terms of the national vote, the most likely outcome is a fairly sizable Democratic plurality of about 53.6% of the two-party vote. But seats are what really matter, and here the Democratic advantage shrinks. Indeed, the most likely outcome based on our model is that the Democrats will win a slim majority, receiving in the neighborhood of 221 seats versus 214 for the Republicans. Given the sources of uncertainty in our model, we actually predict a distribution of outcomes, in which the Democrats win control of the House 54% of the time.

For now, we note that our model suggests a competitive battle for party control of the House even with a forecast of a seven-point spread (53.6-46.4) in the national vote and despite the comparatively large number of incumbents that are not running (37 Republicans vs 18 Democrats as of this writing). As is well known, the Republican

1

advantage in the translation of the vote to seats is largely due to gerrymandering—mostly "natural," due to the concentration of Democrats in heavily Democratic districts (Chen and Rodden 2013), but reinforced by the Republican control of state governments who redistricted following the 2010 Census. The degree of gerrymandering is arguably not much greater, however, than in 2006 when the Democrats comfortably won the House with 54 percent of the vote.[2] An additional factor is the growing partisan polarization. As polarization has increased, the incumbency advantage has shrunk (Jacobson 2015), which withers the GOP advantage from their incumbent majority. With elections decided more based on partisanship than the candidates, the opportunity diminishes for Democrats to gain by running candidates with a strong personal vote. The same applies to Republicans, of course, though they have the majority to lose.

The Model

As discussed, our forecasting model has two steps. In the first, we forecast the national vote division from two variables – the generic poll result and the party of the president. With this estimate of the national swing, the second step forecasts the winners of 435 House races using separate models for open seats and races with incumbent candidates. At both steps, the forecast takes into account uncertainty about the inputs and their effects on the predicted national swing and district vote. The final product of our simulations is a distribution of probable outcomes of the partisan division of House seats. This yields a probabilistic statement regarding the likelihood of the Democratic Party regaining control of the chamber.

Step 1: Predicting the National Vote

We predict the Democratic Party's share of the two-party vote using our two aforementioned independent variables. The first is the current reading of the generic polls—the frequently-asked poll question regarding preferences on a generic (no candidate names) partisan ballot for Congress.[3] The second is a dummy variable for the party holding the presidency (D=1 if a Democratic president and -1 if a Republican president). It is well-known that voters tend to punish the incumbent president's party during midterm elections, and we have shown, based on past congressional campaigns, that generic polls persistently underestimate the ultimate support for the non-presidential ("out") party (see Bafumi et al 2010a). The tendency to underestimate is greatest early in the election year and recedes as the campaign progresses, as poll respondents increasingly take into account the party of the president and tilt toward the out-party. One interpretation is that voters seek more ideological balance between the president and Congress.

Importantly, predictions from the two variables—the generic poll results and the party of the president—are almost equally accurate regardless of when during the election year the poll results are taken. This implies that the structure of the midterm vote is knowable early on, and the campaign mostly serves to draw the voters toward the out-party. Consider that adding election-year changes in the president's approval rating or economic conditions yields no improvement to prediction (Bafumi et al. 2010a). That said, our forecasts are not perfect, partially due to polling error and the fact that other factors can matter on Election Day.

For our forecast, we measure the Democratic percent of the two-party vote (minus 50%) in reported generic ballot polls conducted by personal interview (no internet or robotic polls).[4] With generic polls measured 127 to 156 days in advance of the election, the vote forecasting equation for the 18 midterm elections between 1946 and 2014 is:

National Dem. Vote% = 0.37 % Dem. Polls% − 2.13 Presidential Party + e,   (1)

where the vote and poll variables are measured as deviations from 50%, and Presidential Party = 1 if a Democratic president and −1 if a Republican president. The intercept is suppressed (set to zero) on the tested assumption that polls historically are not biased toward either party.[5] The coefficients for both independent variables are statistically significant at the .001 level. The adjusted R^2 for the equation is 0.76 and the root mean squared error (RMSE) is 2.06.

The pooled generic polls conducted in the 30-day window through June show the Democrats with 54.1% of the two-party vote share. A sizable lead in the polls at this point projects to a significant, if smaller, vote share for the Democrats, even as the out-party. This is because the gain from electorate's tendency to gravitate further toward the "out" party over the midterm year is outweighed by the decline in poll leads.

Our specific forecast is that the Democrats will win 53.6% of the two-party vote and the Republicans the remaining 46.4%. Based on the forecast error of this prediction, the 95% confidence band is a range from 49.4% to 57.8% Democratic.[6] In other words, the Republicans are almost certain to lose the popular vote in 2018. To take into account the uncertainty in our prediction when forecasting, we simulate the vote in 3,000

"elections" based on the forecast error. This yields a probability density as a distribution around the forecast of the national vote.

Step 2: Predicting Seats

We next need to determine how the national vote will impact the number of actual seats the parties win. For each of 3,000 values of the simulated the national vote, we also simulate the outcome in the 435 congressional districts. The simulated vote (V_{jk}) in district k is a function of the stochastic simulation, j, of the national vote plus the local conditions:

$$V_{jk} = N_j + L_{jk}. \qquad (2)$$

The national vote in the jth simulation (N_j) consists of the specific point prediction ($^N P$) from Equation 1 plus the error (e_j) around that prediction. The local (district) component (L_{jk}) in each district k and simulation j consists of the district's prediction ($^D P_k$) based on equations 4 and 5 below plus the error (u_{jk}) reflecting the uncertainty about the prediction. Substituting these components into equation 2 yields:

$$V_{jk} = {}^N P + {}^D P_k + e_j + u_{jk}. \qquad (3)$$

If there is a major party candidate running unopposed in 2016, we assign the seat to the solo candidate's party, even if contested in 2018. For the 343 districts contested in 2016 and presumed to be contested in 2018, we estimate the change in the mean district vote from the difference between the 2016 national vote and our projection of the 2018 vote.

Based on our forecast of the national vote, the expected swing of the national vote, 2016-2018 is 4.1%.

Based on our estimates of the mean district vote swing, we simulate the open seat and incumbent-contested elections. For open seats, our template is the equation predicting the 2016 district two-party vote from the district two-party presidential vote in 2016. This equation for each district k is:

$$\text{DemVote\%2016(House)}_k = -1.26 + 0.96 \text{ DemVotePres\%2016(Pres.)}_k + u_k. \qquad (4)$$

For simulations of open seat outcomes in 2018, we apply equation 4. We adjust the intercept by assigning the predicted mean 2016-2018 vote swing to each open seat for 2018 and add in a normally distributed disturbance for each district based on the root mean squared error (RMSE) (3.81) for Equation 4.

For incumbent-contested seats, our template is an equation predicting the district-level House vote in 2016 from the district-level 2016 presidential vote, the lagged House vote from 2014, plus a term for freshman status (-1 = Republican freshman, +1 = Democratic freshman, 0 = veteran). This equation is:

$$\text{DemVote\%2016(House)}_k = 0.50 + 0.66 \text{ DemVote2014 \%(House)}_k$$
$$+ 0.39 \text{DemVote2016\%(Pres)}_k + 1.16 \text{ Frosh}_k + u_k. \qquad (5)$$

In our simulations of the incumbent-contested seats in 2018, we substitute the 2016 congressional vote for the 2014 congressional vote in Equation 3. We adjust the intercept

so that the overall mean Democratic vote gain in twice-contested districts (2016 and 2018) is equal to the estimated 2016-2018 national Democratic vote swing estimated from the generic poll equation. When simulating 2018 incumbent-contested seats, we include a normally distributed disturbance equal based on the RMSE (3.41) from Equation 5.[7]

Forecasting 2018

To forecast the 2018 election, we first generated 3,000 simulations of the national vote based on equation 1. Then, taking each simulated national vote, we simulated the vote in each congressional district using the formulas shown in equations 4 and 5. This is done for all 2018 open seats as well as seats with incumbents in 2018 that were contested in 2016. For each of the 3,000 simulated vote outcomes, we arrived at a projected outcome in terms of the partisan division of 435 congressional districts. Figure 1 displays the resulting distribution, i.e., density, of outcomes. As might be seen from the preponderance of blue bars, the Democrats win a majority of seats in most of the trials, specifically, 54%. On average, they win 221 seats, which would be 27 more than they won in the 2016 election and would allow a bare 7-seat majority. The simulations do yield considerable variation, however, with a 95% confidence interval of 189 to 253. As can be seen from the positive skew in Figure 1, the Democrats win big in a number of simulations, e.g., gaining 50 or more seats above the 2016 result 9% of the time.

<place figure 1 about here>

Forecasting from October Polls

7

Our forecast can be updated as polls change, indeed, on a daily basis. When this forecast is published in October of 2018, polls at that point in time will be more valuable than those using information from early-July, the publication deadline. One cannot simply plug those polls into equation 1 to forecast the national vote, as the relationship between the vote and October polls differs. Here is the equation using generic polls from the final 30 days of past campaigns in the 18 midterm elections between 1946 and 2014:

$$\text{National Dem. Vote\%} = 0.51 \text{ \%Dem. Generic Poll\%} - 1.14 \text{ Presidential Party} + e, \qquad (6)$$

where, again, the vote and poll variables are measured as deviations from 50% and Presidential Party = 1 if a Democratic president and −1 if a Republican president.

The key difference between equations 1 and 6 is that the in-party penalty is much smaller in October, halved from more than two points using polls in June to about a point. In other words, the party of the president increasingly becomes part of voters' expressed electoral preferences, which matter more for the national vote as the election timeline unfolds. (The adjusted R^2 grows slightly, from .76 to .84.) Should the Democratic share in the generic polls remain close to their current level of about 54.1%, the prospects of a Democratic victory will decline. If history is a guide, however, the Democratic share will grow over the election cycle (Bafumi et al. 2010a).

<place figure 2 about here>

Of course, we do not know what the polls in October will show. To help guide readers, we can estimate the expected seat outcome and probability of Democratic control across a range of poll shares. This is shown in Figure 2. Here we can see that the

8

Democrats are favored to take control of the House as long as they have at least 54% of the two-party share in October's generic polls. At this level of support, the expectation is that they would receive 53.2% or more of the popular vote. However, should they hold a 50% poll share at the end of the campaign, the Democrats would have only about a 12% likelihood of taking the House.

Clearly, then, the Republicans hold a structural advantage in 2018 in terms of translating votes to seats, one that actually has been in place for some time owing to their advantage from the geographic tendency of Democrats to cluster geographically. The GOP strengthened this advantage after the 2010 tide, as many of the new members of Congress were able to avoid defeat when the GOP tide receded in 2012 with their newfound incumbency advantage. The GOP was further helped by the partisan gerrymandering of Republican state legislatures after the 2010 Census. Their larger numbers and associated incumbency advantage will continue to aid them in 2018, though it is worth noting that the advantage has weakened with growing party polarization. That polarization also makes it difficult for Democrats to make inroads in Republican-leaning districts.

Discussion

Inevitably, this modeling effort is based on a set of assumptions. Some readers may be surprised that we do not blithely forecast a Democratic wave in terms of seats based on the Democrats' (as of early July) lead in the polls. Could our assumptions be overly Republican-friendly?

We had choices as to how to model the generic ballot margins in early July. Most that we rejected would tilt the results even less favorable to the Democrats, if only slightly. We also could have chosen to incorporate a constant term in the generic ballot equation. That too would have tilted the estimates to be less Democrats-friendly.

The one bias that could seriously distort our results is the failure to accommodate a possible dynamic that is intractable to modeling based on past elections. The model cannot account for partisan asymmetry where newly energized Democratic campaigns in districts that in the past, Republicans had won by a seemingly comfortable margin. A Democratic surge nationally combined with strong Democratic local campaigns could sway a greater number of districts than modeled here to switch the plurality of their vote. This, we believe, is why our 2010 model underestimated the pull of the 2010 Republican wave. Some cold water on this argument, however, is that the growing partisan polarization could cripple candidate efforts to make districts defy their Republican traditions.

Conclusion

This article offers an estimate of the likely distribution of seats in the House following the 2018 election. To provide our estimate, we first forecast the national vote based on the historical relationship between generic polls, the party of the president, and the vote in previous midterm elections. Taking into account the expected vote using June polls and the unique circumstances of 435 House districts, we then simulate the Election Day outcome. The average result of our simulations is a Democratic-controlled House with a narrow 7-seat majority. But this is only an average, with a wide dispersion of

possible outcomes. The model suggests that the Republicans have a 46% chance of retaining control.

This prediction is based on a key assumption that might turn out to be untrue. That assumption is that the error term—the unobserved variance due to candidates and their campaigns—is constant across districts with a mean of zero. The lore of the 2018 campaign is that Democrats are running stronger races than normal in districts that they previously had largely conceded as unwinnable. This intangible factor cannot be incorporated into our model, but could be subject to validation with the arrival of district-level polls later in the campaign. (The polls that currently are available – as well as the swing in special elections – are suggestive.) If a Democratic surge does concentrate in districts that became newly winnable, an extra edge goes to the Democrats. Our results might thus be best interpreted as a lower bound of possible outcomes for the Democratic Party and a best-case for the Republicans in 2018.

References

Bafumi, Joseph, Robert S. Erikson, and Christopher Wlezien. 2008. "Forecasting House Seats from Generic Congressional Polls." In *Elections and Exit Polling*, ed. Wendy Alvey and Fritz Scheuren. New York: Wiley and Sons, 118-123.

Bafumi, Joseph, Robert S. Erikson, and Christopher Wlezien. 2010a. "Balancing, Generic Polls, and Midterm Congressional Elections." *Journal of Politics* 72:705-719.

Bafumi, Joseph, Robert S. Erikson, and Christopher Wlezien. 2010b. "Forecasting House Seats from Generic Congressional Polls: The 2010 Midterm Elections." *PS: Political Science and Politics* 43:633-36.

Bafumi, Joseph, Robert S. Erikson, and Christopher Wlezien. 2014. "National Polls, District Information, and House Seats: Forecasting the 2014 Midterm Election." *PS: Political Science and Politics* 47:775-778.

Chen, Jowei, and Jonathan Rodden. 2013. "Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures." *Quarterly Journal of Political Science* 8:239-269.

Jacobson, Gary C. 2015. "It's Nothing Personal: The Decline of the Incumbency Advantage in US House Elections." *Journal of Politics* 77(3): 861-873.
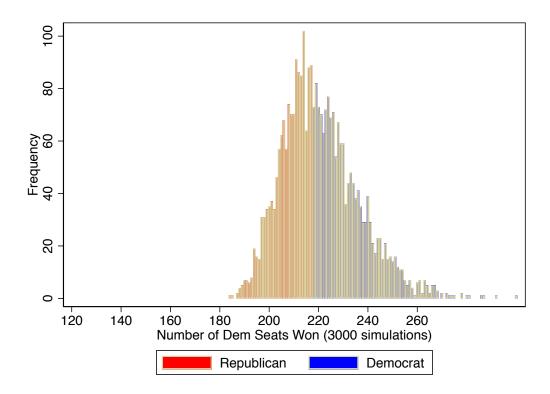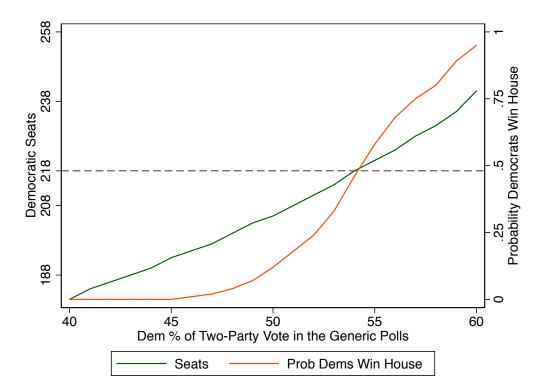
12

Figure 1. Three Thousand Simulations of the 2018 Election

Figure 2. Simulations of the 2018 Election Outcome, Conditional on the Generic

Polls during the 30 Days before Election Day

---

[1] Using an earlier version of the model presented here, we forecasted a 32-seat gain by

the Democrats in 2006, close to the 30 seats they actually gained (Bafumi et al 2008). We

did less well in 2010, forecasting a 50-seat gain by the Republicans, which was 13 short

than what they realized on Election Day (Bafumi et al 2010b). In 2014 we forecasted the

Republicans would win 248 seats, overshooting the final outcome by only 1 seat (Bafumi

et al 2014).

[2] As evidence, we can compare the 2008 Obama vote in districts created following the 2000 and the 2010 Census. The post-2000 redistricting only fractionally decreased the proportion of districts where Obama won a plurality of the 2008 presidential vote.

[3] Respondents typically are asked for which party they plan to vote (or whom they want to win) "if the election were being held today," although there is variation in the wording used. For instance, some organizations use the wording "looking ahead to the Congressional elections in November" while others use "thinking about the next election for U.S. Congress." We do not attempt to account for any effects of these differences. Also, polls may be asked of adults, registered voters or likely voters. Separate analysis suggests that these differences matter little in recent times so we do not account for them.

[4] This restriction is not intended as a statement about the relative accuracy of different kinds of polls, and is done to ensure procedural continuity. We use polls listed by 538.com.

[5] Not including an intercept is equivalent to predicting the presidential party vote from the generic polls but without a dummy variable for which party holds the presidency. When an intercept is included in the equation, it is slightly negative and statistically insignificant. Including an intercept would produce a slightly more Republican-friendly prediction of the vote.

[6] The forecast error (2.17) takes into account both the root mean squared error of the equation and the uncertainty about the coefficients.

[7] The following details are our solutions to some minor issues in forecasting district-level outcomes:

15

Special elections: We treat 2018 districts with bi-elections (also known as special elections) between 2014 and 2016 as incumbent races. We use their bi-election vote outcomes but adjust them to reflect the likely outcome had the election been held in 2016. First we regress the 2016 vote in these districts on the 2016 presidential vote and generate predicted values from the regression. The average difference between the predicted vote in 2016 and the vote in the bi-election is used to adjust the bi-election outcome.

Redistricting: Pennsylvanian redistricted its entire state delegation due to a Pennsylvania Supreme Court mandate. With district boundaries disrupted, we treated all Pennsylvania seats as if open seats. For the two Pennsylvania districts with no contest in 2016, we proxied the 2016 "vote" by an estimate based on the district's presidential voting. For the one Pennsylvania seat that appears to be uncontested in 2018, we treat as if competitive. These steps ensure a block of seats to be treated as competitive in both elections, with an average vote swing equal to the expected national swing.