Harvard Data Science Review • Issue 2.4, Fall 2020

An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election

Merlin Heidemanns, Andrew Gelman[1], G. Elliott Morris[2]

[1]Department of Statistics and Department of Political Science, Columbia University, [2]The Economist

Published on: Oct 27, 2020

DOI: 10.1162/99608f92.fc62f1e1

ABSTRACT

During modern general election cycles, information to forecast the electoral outcome is plentiful. So-called fundamentals like economic growth provide information early in the cycle. Trial-heat polls become informative closer to Election Day. Our model builds on (2013) and is implemented in Stan Stan Development Team, 2020). We improve on the estimation of state-level trends, the internal consistency of different predictions at the state and national level, and provide an adjustment for differential nonresponse bias across the cycle. The model forecast a Democratic win with probability in the 80–90% range during most of the 2020 U.S. presidential election campaign, conditional on the two major candidates staying in the race, no major third-party challenges, and no unprecedented challenges with turnout or vote counting.

Keywords: poll aggregation, election, forecasting, fundamentals, Bayesian

Media Summary

We forecast vote intentions for Election Day by combining fundamentals and trial-heat polls. Fundamentals such as economic growth and presidential approval are highly predictive early in the electoral cycle. Trial-heat polls become more predictive as we approach Election Day. Our model focuses on the internal consistency of predictions at different levels, e.g., how likely is a win in a particular state if the candidate leads with a certain margin at the national level. Our model also includes an adjustment for differential nonresponse of Democratic and Republican voters to correct for polling artifacts that in 2016 led to short-term shifts in poll numbers as more or fewer partisan supporters participated throughout the cycle. The model forecast a Democratic win with probability in the 80–90% range during most of the 2020 U.S. presidential election campaign, conditional on the two major candidates staying in the race, no major third-party challenges, and no unprecedented challenges with turnout or vote counting.

---

1. Introduction

We constructed an election forecasting model for *The Economist* that builds on Linzer's (2013) dynamic Bayesian forecasting model and provides an election day forecast by partially pooling two separate predictions: (1) a forecast based on historically relevant economic and political factors such as personal income growth, presidential approval, and incumbency; and (2) information from state and national polls during the election season. The two sources of information are combined using a time-series model for state and national opinion. Our model also accounts for some aspects of non-sampling errors in polling. The model is fit using the open-source statistics packages R and Stan (R Core Team, 2020; Stan Development Team, 2020) and is updated every day with new polls. The forecast is available and

is conveyed with several dynamic graphics displaying data and predictions of electoral and popular votes at the national and state level. See also: a description of the model-building process and all code. This paper provides a more formal description of our statistical model than is available elsewhere. While more thorough, the model code posted on the GitHub page should be seen as the authoritative source of how the model works, following the principle that 'code never lies.' We also describe how our model builds on the existing literature surrounding Bayesian election forecasts, namely in accounting for political polarization and differential partisan nonresponse in the polls.

2. Polls

We include polls at the national and state level and take each poll to be an estimate of that day's average support for the Democratic and Republican candidates for president (ignoring respondents who express no opinion or support other candidates), with modeled bias and variance. Our goal is to estimate national and state-level trends in support for the candidates. We start by considering how we model individual polls before discussing how individual components are modeled.

For each poll i the number of respondents indicating their support for the Democratic candidate is given by y_i with n_i being the total number of respondents supporting one of the major party candidates. We start with the binomial sampling model:

$$y_i \sim \text{Binomial}(\theta_i, n_i).$$

We model θ_i conditional on whether poll i is at the state or national level. We model state polls in state $s[i]$ at time $t[i]$ as

$$\theta_i = \text{logit}^{-1}\left(\mu^b_{s[i],t[i]} + \alpha_i + \zeta_i^{\text{state}} + \xi_{s[i]}\right)$$

and national polls as

$$\theta_i = \text{logit}^{-1}\left(\sum_{s=1}^{S} w_s \mu^b_{s,t[i]} + \alpha_i + \zeta_i^{\text{national}} + \sum_{s=1}^{S} w_s \xi_s\right).$$

In our notation, we are using superscripts as names and subscripts as indexes; for example, in the expression, $\mu^b_{s,t}$, the parameter is named μ^b and it is indexed by state s and date t.

The most important term in the above formulas is $\mu^b_{s,t}$, which represents the underlying support for the Democrat in state s at time t. For national polls, these are summed over the $S = 51$ states (including Washington, D.C.) with weights w_s that sum to 1 and which are proportional to the number of votes in the state in the previous election. Our model does not account for third parties and would need to be expanded to apply to an election with large third-party vote shares.

The other terms in the above model, α, ζ, ξ, represent different sources of bias arising from the well known fact that opinion polls are not actually random samples of the voting population: Such a feat would be impossible given high nonresponse rates of modern surveys.

We expand the shared bias term in the state and national poll models as,

$$\alpha_i = \mu^c_{p[i]} + \mu^r_{r[i]} + \mu^m_{m[i]} + z\epsilon_{t[i]},$$

including house effects μ^c, polling population effects μ^r, polling mode effects μ^m, an adjustment trend term for nonresponse bias ϵ and an indicator z equal to 1 if the pollster does not adjust for partisanship and 0 otherwise, measurement error ζ, and state-level error ξ.

In the following, we cannot disclose most of the specific information on the hyperparameters, as well as the fundamentals-based forecast, as these are proprietary to *The Economist*. Hyperparameter choices for past election cycles can be gleaned from the code provided in the GitHub repository for those elections.

2.1. State-level trends

We share information across states contemporaneously and over time. We accomplish this by treating the development of state-level public opinion as a correlated random walk for which we have prior information from the fundamentals-based prediction at $t = T$ (Election Day).

The random walk component connects days for which we have polls and interpolates for days without polls in between. Our uncertainty for days without data is a function of our encoded prior model of how much public opinion can change from day to day as well as the polling data from other states. We cross-validated this value on the 2008, 2012, and 2016 elections and the first months of 2020.

The correlation in the random walk imposes our assumption on the estimates that in the absence of data similar states will move in similar ways, i.e., if we have polls for Washington but not for Oregon, then the daily trend for Oregon will look similar to the trend in Washington with added uncertainty. We set the details of this correlation matrix so as to obtain reasonable results for national and state-level swings. We estimated the correlation matrix Σ from past election results and other relevant state-level predictors such as education. We set off-diagonal elements smaller than zero to zero and then scale the matrix to achieve the desired degree of day-to-day change.

Then, the process takes the form

$$\mu^b_t | \mu^b_{t-1} \sim \mathrm{MVN}\left(\mu^b_{t-1}, \Sigma^b\right),$$

where μ^b_T is the estimate from the fundamentals model.

2.2. House, population, and mode effects

We include adjustment terms for pollsters, polled population (e.g. likely voters and registered voters), and polling mode (e.g. live caller or online). We know that pollsters can favor either party, and that the same applies to the polled population and how the populations are being polled. Our priors for the parameters are centered at 0 with different standard deviations:[1]

$$\begin{aligned}\mu^c &\sim \text{Normal}(0, \sigma^c)\\ \mu^r &\sim \text{Normal}(0, \sigma^r)\\ \mu^m &\sim \text{Normal}(0, \sigma^m).\end{aligned}$$

This implies that we are estimating the deviation given this year's polling data, thus not using potential information about the quality of the pollsters or the reliability of likely voter adjustments from past data, out of concern that these may not provide reliable indications for the current election.

2.3. Partisan nonresponse adjustment term

Poll-aggregation election forecasts performed poorly in 2016, a problem that can be attributed to polls in key midwestern states that did not appropriately adjust for nonresponse (Gelman & Azari, 2017). This adjustment is represented by an autoregressive process to allow the party adjustment to vary over time at the national level; see Gelman et al. (2016). If a pollster does not adjust for the partisan composition of their sample, shifts in support can reflect a changing sample composition. We rely on the difference between adjusting and non-adjusting polls to estimate the extent to which non-adjusting polls are biased:

$$\xi_1 \,|\, \rho, \sigma^\xi \sim \text{Normal}\left(0, \frac{1}{\sqrt{1-\rho^2}}\,\sigma^\xi\right)$$

$$\xi_t \,|\, \xi_{t-1}, \rho, \sigma^\xi \sim \text{Normal}\left(\rho\xi_{t-1}, \sigma^\xi\right) \text{ for } t = 2, \ldots, T$$

$$\rho \sim \text{Nnormal}\left(0.7, 0.1\right).$$

2.4. Measurement error

We add additional uncertainty to each poll where the scale varies based on whether it is a state or a national poll. These terms are unidentifiable, i.e., given N polls, there are N independently and

identically distributed terms, but we adjust our uncertainty in the poll estimates, for example for poll specific design effects or deviations from truly random samples. That is, we assume

$$\zeta_i^{\text{national}} \sim \text{Normal}(0, \sigma^{\text{national}})$$
$$\zeta_i^{\text{state}} \sim \text{Normal}(0, \sigma^{\text{state}}).$$

2.5. Correlated state errors

We include state and national level polling error terms, which allows for unmodeled measurement error for each poll beyond the stated margin of error (Shirani-Mehr et al., 2018). We treat state level polling error terms as correlated across states with a scaled version of the same correlation matrix we use for changes in underlying opinions across states:

$$\xi \sim \text{MVN}(0, \Sigma^{\xi}).$$

Furthermore, we consider the same relationship as with μ^b for the national polls where we include the weighted sum of the state-level error term as the national level error. The weights again are the forecast state-level shares, and for these we simply have used voter turnout in the previous presidential election.

3. Fundamentals

The fundamentals-based model combines the previous electoral outcome with economic and political factors, based on the "time for change" model of Abramowitz (2008), but we add an interaction term between economic growth and the share of swing voters in the electorate, as measured by the American National Election Study, to allow for dampening effects of economics in polarized elections. We predict the incumbent vote share by state in previous elections using a regularized linear model and predict the incumbent vote share in 2020 with the parameter estimates. We set the prior for μ_b on Election Day to the fundamentals-based prediction.

4. Putting the Pieces Together

We combine the two forecasts by using the fundamentals-based prediction as the prior for Election Day. The random walk prior on μ^b can be visualized as going backward in time from Election Day to the current day of polling. Thus, the model updates the prior for Election Day by the poll-based forecast for Election Day. Figure 1 shows the model fit for 2016. As with other forecasts, our model overstates the strength of Hillary Clinton in key midwestern states (see Michigan in the graph) because of failures in the state polls, but its hierarchical model with multiple error terms allows the model to avoid the over-certainty that could arise from simple poll averaging.

The current prediction for 2020 can be found on the website of *The Economist*.

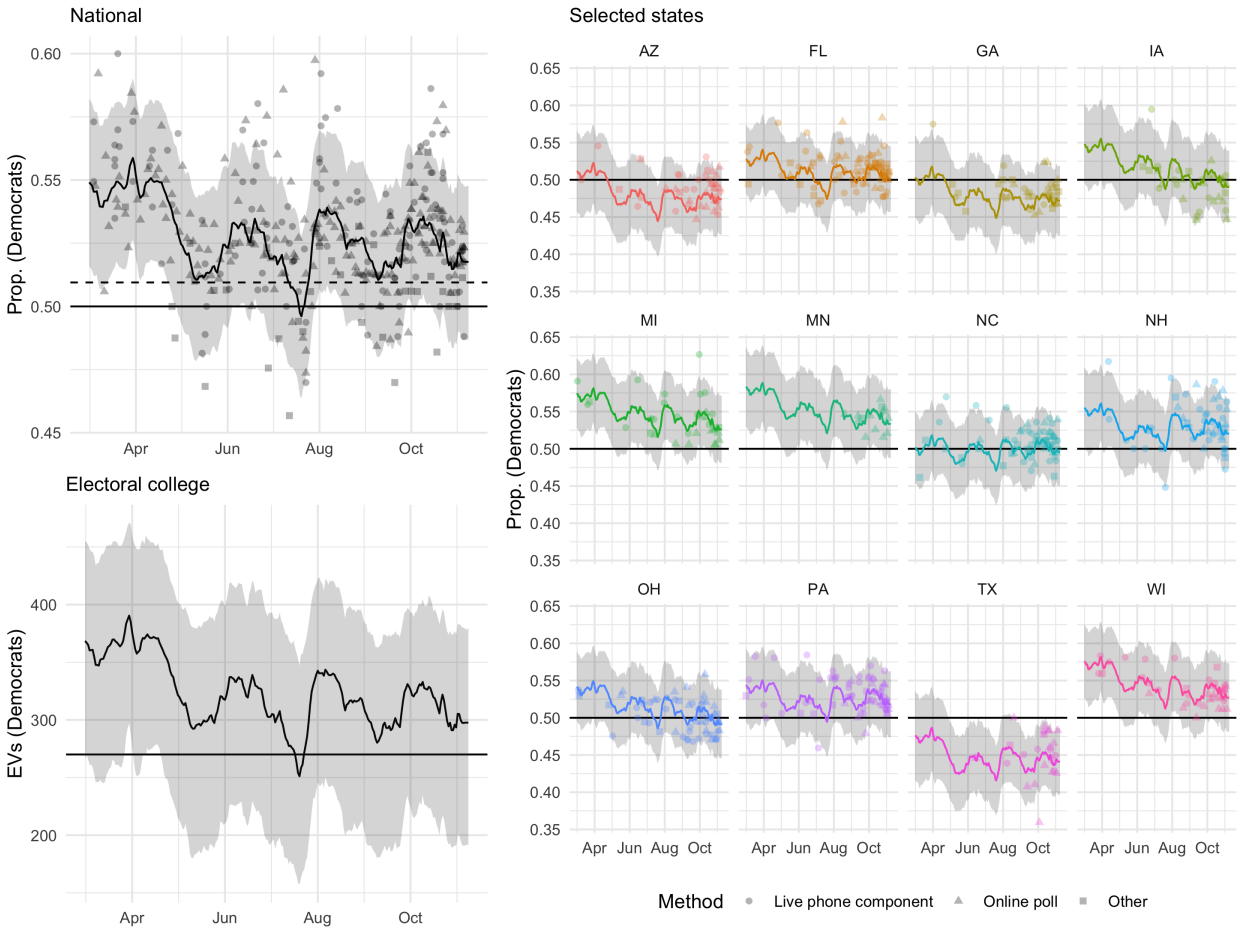

Figure 1. Some summaries of the model, as fit retrospectively to using state and national polls from 2016. These graphs illustrate that our data and model are fitting national as well as separate state trends,

5. Calibration, Uncertainty, and What Is Forecast

The model estimates a large number of parameters with a relatively small number of polls. To estimate the parameters, we must supply relevant prior information. These pertain to our chosen priors, the predictors in the fundamentals-based forecast, and the construction of the covariance matrix that shares information across states. Model results are thus sensitive to these choices. In making these choices, we want to avoid unwarranted precision (e.g., a prediction that Biden will win Florida and with 95% probability that his share is between 51% to 52%) and unwarranted uncertainty (e.g., Biden's share will be between 40% and 60% with 95% probability). As part of the Bayesian workflow, we started with values that we deemed reasonable a priori such as a 3% polling error for each poll based on historical data, but also evaluated the model output to determine whether the model gave sensible results. For example, based on our knowledge of the electorate in Florida, believing that the Republican candidate could win the state with 60% of the vote would be deemed unreasonable given increased partisanship and its status as a swing state.

We model vote intentions rather than the electoral outcome. Our prediction for the Electoral College translates directly from our forecast vote intentions, i.e., winning Maine in 70 out of 100 simulations on average. We for example do not consider the effect of COVID-19 on turnout or absentee ballot rejection rates that might differentially penalize the Democratic candidate.

Finally, there will always be the potential for further checking and improvement of the model. In the four months between the initial release of our model in June, 2020, and the time of this writing, we have discovered or have been informed of several questionable predictions from our model as revealed in its fit to the current and previous elections, leading us back to our code, where we discovered some bugs and questionable modeling choices. One advantage of openness—our code is available on Github, updated predictions appear on *The Economist* site every day, and we have had several free-flowing discussions on our blog (examples one and two)—is that we have engaged a broad community of active readers who have helped us poke at our predictions in many different ways. The model described in the present article should be thought of not as a final product but rather as a step along a continuing path. Gelman et al. (2020) provide further discussion of the choices we and others have made in modeling and communicating election forecasts.

6. Conclusion

Forecasting an election is complex and can be framed as even more so in an unfamiliar environment. Potentially widespread absentee voting may change both turnout as well as the share of the population whose votes are counted. Economic shocks usually reflect negatively on the incumbent but may not if induced due to a global pandemic. Pollsters may be more actively partisan than they were in previous elections. Overall, our model accounts for a variety of factors and treads carefully when it comes to choosing between overconfidence and expressed helplessness due to the plethora of unfamiliar events. That is, we focus on the factors we can credibly model but also believe that this election, at least with respect to modeling vote intentions, is not fundamentally different from the previous elections we used to calibrate it.

Disclosure Statement

The authors have nothing to disclose.

References

Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science & Politics*, *41*, 691–695.

Gelman, A., & Azari, J. (2017). 19 things we learned from the 2016 election (with discussion). *Statistics and Public Policy*, *4*, 1–10.

Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science*, *11*, 103–130.

Gelman, A., Hullman, J., Wlezien, C., & Morris, G. E. (2020). Information, incentives, and goals in election forecasts. *Judgment and Decision Making*, *15*, 863–880.

Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, *108*, 124–134.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*, 607–614.

Stan Development Team. (2020). *Stan modeling language*. http://mc-stan.org/

Footnotes

1. We are following the convention of Stan, where the univariate normal distribution is parameterized by its mean and standard deviation, and the multivariate normal distribution is parameterized by its mean vector and covariance matrix. ↩

Citations

1. Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, *108*, 124–134. ↩

2. Team, S. D. (2020). *Stan modeling language*. http://mc-stan.org/ ↩

3. Team, R. C. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/ ↩

4. Gelman, A., & Azari, J. (2017). 19 things we learned from the 2016 election (with discussion). *Statistics and Public Policy*, *4*, 1–10. ↩

5. Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science*, *11*, 103–130. ↩

6. Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*, 607–614. ↩

7. Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science & Politics*, *41*, 691–695. ↩

8. Gelman, A., Hullman, J., Wlezien, C., & Morris, G. E. (2020). Information, incentives, and goals in election forecasts. *Judgment and Decision Making*, *15*, 863–880. ↩