Prediction of the 2018 Midterm Elections

Jiafeng Chen* Joon Hyuk Yang†

November 6, 2018

1 Introduction

We present a statistical model and a machine learning method to perform prediction for the 2018 elections of the U.S. House of Representatives. The main challenge of such a prediction problem is that the historical panel of election data is short—both in terms of years with adequate non-missing covariates (2010–2016) and in terms of observations (≤ 435 per year). The scant supply of historical data creates a generalization problem. All elections in 2010–2016 are considered Republican wave elections,[1] and as such, models trained on this data may reasonably favor Republicans. This problem could be mitigated if we condition on more data and avoid overfitting. However, the low number of observations becomes extremely problematic, as $p > N$ problems are quickly encountered—especially with complex basis transforms over the data. Furthermore, the changing political landscape of 2018 makes real-time polling data quite valuable, but with the lack of historical polling data, we cannot simply incorporate polling as an additional feature.

We devise a Bayesian approach that seeks to mitigate these problems. In our approach, the training data (data before 2018) is used to construct a prior, which is then updated with the polling data from 2018. Our model resembles what Nate Silver calls FiveThirtyEight's Classic House Forecast Model,[2] where our prior estimation is analagous to Silver's "fundamentals", while our polling updating mirrors the "simulating the election" in Silver's model. Our approach is fully compatible with additional features or polls and is flexibile enough to admit several extensions. Moreover, thanks to a few simplifying approximations, our approach is extremely computationally tractable while remaining statistically faithful.

This report is organized as follows: Section 2 discusses our statistical model and machine learning method, including parameterizations, estimation strategies and the considerations therein. Section 3 discusses various datasets we use and preprocessing applied. Section 4 displays various updated predictions and provides a brief discussion of potential shortcomings and alternative strategies. Section 5 concludes with potential areas of improvement and further exploration.

*Harvard College, jiafengchen@college.harvard.edu
†Harvard College, joonhyukyang@college.harvard.edu
[1] Maybe save for 2012, which is a presidential election that Democrats won.
[2] https://fivethirtyeight.com/features/2018-house-forecast-methodology/

2 Statistical model and prediction function

2.1 Overview

Let

$$Y_i = \frac{\text{Republican\%}}{\text{Republican\%} + \text{Democrat\%}}$$

be the proportion of Republican vote share between the major parties in district i, where i is indexed by the triple (state, district number, year of the race). Let $x_i \in \mathbb{R}^p$ be a list of features (after suitable basis transformations) for the district.[3]

A machine learning method can generate a prediction function $\widehat{f}_{\mathcal{T}}$, where $\mathcal{T}$ is the collection of Y_i, x_i over the training set. However, the distribution of the test set, $Y_i \mid x_i$ for those districts i in 2018 may be substantially different from that of $Y_j \mid x_j$ for j in the training set. Thus $\widehat{f}_{\mathcal{T}}$ may incur large generalization errors, since we are attempting to generalize the prediction function to a potentially *different* distribution of the data.[4] Fortunately, we have an additional piece of information, namely polling data, that we may scrape from polling aggregators like FiveThirtyEight.[5] The lack of historical polling data means that we cannot simply include them in the features x_i. Therefore, to incorporate polling data, we perform a two-step Bayesian procedure. In the first step, we estimate a prior $Y \sim p_{\theta_0,X}$, where parameters θ_0 is estimated from the training data. In the second step, we assume that polling data, Z, comes from some conditional distribution $Z \mid Y \sim p_\eta(z \mid y)$ for some known or estimated parameters η. We form our final prediction by computing the posterior distribution $Y \mid Z$. We now discuss how we parameterize $p_{\theta,X}$ and p_η.

2.2 Parameterizations

For the prior, $p_{\theta,X}(y)$, we assume a linear probability model,

$$Y \sim \mathcal{N}(\underbrace{X\beta_0}_{\mu_0}, \Sigma_0) \tag{1}$$

(1) is misspecified since Y_i is between 0 and 1, but the Normal distribution is supported on $\mathbb{R}$. A properly specified generalized linear model for Y_i is a Beta linear model (See Grün, Kosmidis, and Zeileis, 2011, for an overview). We have a few reasons for preferring (1) instead. First, we note that for large values of a, b in Beta(a, b), the distribution Beta(a, b)

[3] We are assuming that each competitive district has a Republican and a Democratic candidate, and that third-parties never win elections. The assumption is largely true, as third parties are not favored to win any house elections in 2018 (albeit two Senators, Bernie Sanders and Angus King, are independents). The assumption is sometimes wrong, as in the case of California–8. Results in Appendix A predict that Donnelly (R) will win, but in fact California–8 has two *Republican* candidates running against each other (and no Democratic candidate), and Donnelly is widely projected to lose. The reason that our model err is that we only put Donnelly as the single Republican candidate with no Democratic opponent.

[4] Note that this problem is mitigated by increased data availability—for instance, if we had obtained historical polling data, which is difficult to obtain—but we incur a different problem ($p > N$) of high-dimensional inference in that case.

[5] https://projects.fivethirtyeight.com/polls/

is approximately Normal. Second, we note that

$$Y_i = \sum_{j=1}^{M} \frac{\mathbb{1}(\text{vote}_j \in \{D, R\})}{\sum_{j'=1}^{M} \mathbb{1}(\text{vote}_{j'} \in \{D, R\})} \mathbb{1}(\text{vote}_j = R),$$

for a district with M voters and parties D, R. Y_i should have an approximately Normal distribution, by the Central Limit Theorem, since it is a type of empirical average. Third, computational methods for the Beta regression model—especially those that deal effectively with high-dimensional covariates—is much less readily available than those for the Normal linear model. We note additionally in Figure 5 that the fitted residuals are approximately Normally distributed and in Figure 6 that they are homoskedastic, and we note that Normal distribution eases computation for Bayesian updating in our next step. For (1), we need to estimate $\theta_0 = (\beta_0, \Sigma_0)$.

For p_η, we assume that conditional on Y, the polls Z_j are independently distributed. We can treat an observation of a poll z_j as generated by asking n_j individuals a question ("Are you voting for the Republican candidate"), with $z_j \in [0, 1]$ respondents responding affirmatively.[6] Assume that

$$Z_j \mid Y \sim \mathcal{N}(a_j^T Y, \sigma_{Z_j}^2), \text{ for } a_j, \sigma_{Z_j}^2 \text{ that do not depend on } Y. \tag{2}$$

We use the Normal distribution in (2) because (1) The Normal distribution works well computationally with our Normal prior[7] and (2) polling proportions are empirical averages, which tend to be Normally distributed by the Central Limit Theorem.

We use two types of polls in our implementation. First is a national *generic ballot* poll, which simply asks the respondent which party she would vote for. For such Z_j, we assume $a_j = n^{-1}\mathbf{1}$, treating the mean of such a poll as aggregated over all contested districts. Second is a district-wide poll, and we use $a_j = e_i$ for a poll in district i, where e_i is the i-th standard basis vector. For (2), we need to estimate $\sigma_{Z_j}^2$ for both types of polls.

We now discuss our approaches to estimation and justify discretionary choices made therein.

2.3 Estimation

2.3.1 Estimation for θ_0

We first discuss estimation for $\theta_0 = (\beta_0, \Sigma_0)$. Let

$$\widehat{\beta}_0 = \arg\min_{\beta} \sum_{i=1}^{|\mathcal{T}|} w_i (y_i - x_i^T \beta)^2 + \lambda \left(\alpha \|\beta\|_1 + (1 - \alpha) \|\beta\|_2^2 \right) \qquad \text{(Estimation for } \beta_0\text{)}$$

[6] Again, this model assumes that the only parties running are Republicans and Democrats, and that there are no undecided individuals. When working with data, we normalize the Republican percentage by the sum of the Republican and Democratic percentages.

[7] This is assuming that the Normal likelihood has variance unaffected by the prior—which requires a further justification that we discuss in Section 2.3.

be fitted with an elastic net regularizer over the training data with weights w_i, where (α, λ) is chosen via 10-fold cross-validation. The weights we chose is uniform $w_i = 1$; however, in the spirit of efficient estimation as in weighted least squares, it may be desirable to choose the weights to the proportional to total votes cast, since we should expect that variance of the measurement Y_i is inversely proportional to the number of votes cast, as Y_i is an empirical average. In numerical experimentation, we found that the weights do not affect results much, and we keep $w_i = 1$ for simplicity.

It is crucial, for the purpose of uncertainty quantification, that we perform good variance estimation. Let $S_i = \{j : \mathsf{state}(j) = \mathsf{state}(i)\}$ be those districts in the same state as i (across time),

$$\widehat{\Sigma}_{ii} = \frac{1}{|S_i|} \sum_{j \in S_i} \left(y_j - x_j^T \widehat{\beta}_0 \right)^2 \qquad \text{(Estimation for variance)}$$

be an estimate for variance. Here we estimate variance by pooling over the state of a district, as states boundaries, unlike district boundaries, do not change over time. Alternatively, we could estimate variance by letting S_i be those districts that share the same state and district identifier (which may not be the same geographical area due to redistricting), or with the k nearest geographical neighbors of district i (in the spirit of kNN).

Optionally, we also estimate off-diagonal elements of Σ. If district boundaries do not change over time, then we can easily estimate off-diagonal entries of Σ as well, since we can use the variation over time to calculate the empirical analogue of covariance:

$$\widehat{\Sigma}_{ij} = \frac{1}{T} \sum_{t=1}^{T} \left(y_{it} - x_{it}^T \widehat{\beta}_0 \right) \left(y_{jt} - x_{jt}^T \widehat{\beta}_0 \right) \qquad \text{(Estimation for covariance)}$$

This approach is no longer valid when district boundaries change over time. Nonetheless, we can still calculate $\widehat{\Sigma}_{ij}$, ignoring district boundary changes. Note that (Estimation for variance) is inconsistent with (Estimation for covariance), and filling in estimates fromthe covariance estimation directly into Σ would typically generate non-positive-definite matrices.[8] As a result, we estimate the *correlation* $\widehat{\rho}_{ij}$ similar to (Estimation for covariance),[9] and use the estimate

$$\widehat{\Sigma}_{ij} = \widehat{\rho}_{ij} \sqrt{\widehat{\Sigma}_{ii} \widehat{\Sigma}_{jj}} \qquad \text{(Estimation for covariance with correlation)}$$

which ensures that the resulting Σ_0 is positive definite. Figures 6 and 7 displays the Bayesian updating for both specifications, Σ_0 diagonal and Σ_0 fully general. They display signifcant differences—we see that the error correlation affects the uncertainty in prior in terms of Republican seats greatly.

Now we discuss a few caveats regarding variance estimation. In principle, the variance we estimate should also take into account the uncertainty in $\widehat{\beta}_0$. However, there is little consensus in uncertainty quantification for methods like elastic net and LASSO (Kyung et al., 2010), which remains an active research area. The fitting (along with cross-validation) is too

[8] Alternatively, of course, we could discard the smoother in (Estimation for variance) and just estimate variance as in (Estimation for covariance). We think the smoother is a more principled approach that takes into account redistricting. The approach in (Estimation for covariance) is a second-best.

[9] We fill undefined values of $\widehat{\rho}_{ij}$ with zero.

expensive to bootstrap either, and we ignore the uncertainty in $\widehat{\beta}_0$ for simplicity. Moreover, (Estimation for variance) is likely an underestimate of variance—in linear regression, for example, we inflate the naive variance estimate by $n/(n-p)$, where p is the degrees of freedom in a linear regression. The analogous inflation in our setting is with the degree of freedom in the elastic net (Zou et al., 2007)—the estimate of the degree of freedom is about 100 for a dataset with $N = 1500$; thus the inflation factor is about $\frac{1500}{1500-100} \approx 1.07$ for variance, which is sufficiently small to ignore. An alternative would be to estimate variance on a hold-out set, which is unbiased if the data is i.i.d.; we do not hold out for the sake of increasing training data.

As a notational matter, we now suppress the hat symbol on β_0, Σ_0.

2.3.2 Estimation for polling variance

We now turn to estimation of variance in the likelihood of polls, $\sigma^2_{Z_j}$ in Equation (2). The main problem in variance estimation is to account for the large between-poll variance unpredicted by a simple model. Suppose poll outcomes are indepedently $n_j^{-1}\,\mathrm{Bin}(n_j, Y_i)$ conditional on Y_i, then the (conditional) standard deviation is as low as 1.58 percentage points for a poll with 1000 respondents. However, polls of this size display much more variation than predicted by a simple independent Binomial model. Therefore, we assume that there is some structural between-poll variance and attempt to estimate it from data. More precisely, we view the poll as

$$Z_j \mid Y \sim n_j^{-1}\,\mathrm{Bin}(p(Y), n_j) + \epsilon_j,\ \epsilon_j \perp\!\!\!\perp Y,$$

where ϵ_j represents a between-poll variation, and the variance $n_j^{-1}\,\mathrm{Bin}(p(Y), n_j)$ is a within-poll variance resulting from sampling error.

For both types of polls, there is a within poll variance generated by sampling, and a between-poll variance. The within poll variance is $n_j^{-1}\mathbb{E}[Z_j \mid Y](1 - \mathbb{E}[Z_j \mid Y]) \le (4n_j)^{-1}$, which in general depends on Y. For computational simplicity, we use the upper bound $\frac{1}{4n_j}$ as a proxy for the within-poll part of variance, since it does not depend on Y_i in order to take advantage of Bayesian conjugacy; the approximation is good for districts with Y_i close to $1/2$, which is precisely those polls that we especially care about. Full, proper posterior condition can still be achieved if we use the more precise estimate, $Y_i(1-Y_i)/n_j$, via an MCMC sampler.

For a generic ballot poll Z_j, we leverage the FiveThirtyEight polling average estimates since FiveThirtyEight provides a 90% confidence interval. We use the (most recent) variance that implies the range of 90% confidence computed by FiveThirtyEight. The FiveThirtyEight estimates are via kernel-weighted local polynomial smoothing,[10] whose confidence comes from (we conjecture) bootstrapping. We directly use FiveThirtyEight's estimates, as the between-poll part of the variance for generic-ballot polls, for convenience.

For a district-wide polls $Z^i_j, j = 1, \ldots, J_i$, the variance comes from two components. The first component comes from the sampling error of the poll itself, which is the Binomial variance $\frac{Y_i(1-Y_i)}{n_j} \le \frac{1}{4n_j}$. The second component comes from a between-poll variance, which

[10]The link on footnote 2 in FiveThirtyEight's methodology links to the Stata manual for kernel-weighted local polynomial smoothing.

we estimate by taking the empirical variance of poll results in a district.[11]

To summarize, we use

$$\sigma_{Z_j}^2 = \frac{1}{4n_j} + \sigma_{538}^2 \qquad \text{(Estimation for generic ballot poll variance)}$$

$$\sigma_{Z_j}^2 = \frac{1}{4n_j} + \widehat{\sigma}_{i,\text{poll}}^2 \qquad \text{(Estimation for district poll variance)}$$

where σ_{538}^2 is the variance implied by FiveThirtyEight's 90% confidence interval, and $\widehat{\sigma}_{i,\text{poll}}^2$ is the empirical variance of polling results for district i estimated by taking the empirical variance within district i.

2.4 Putting it all together: updating prior

From estimating θ_0, for the competitive races in 2018, we obtain a prior estimate $Y \sim \mathcal{N}(\mu_0, \Sigma_0)$, where $\mu_0 = X\beta_0$. The polls form a vector $Z \mid Y \sim \mathcal{N}(AY, \Sigma_Z)$ for A, Σ_Z compatible with our assumptions.

Therefore, our posterior is

$$Y \mid Z \sim \mathcal{N}\left(\mu_0 + \Sigma_{ZY}\Sigma_Z^{-1}(Z - AY), \Sigma_Z - \Sigma_{ZY}\Sigma_0^{-1}\Sigma_{YZ}\right), \tag{3}$$

by the usual Normal conditioning procedure, with $\Sigma_{ZY} = \text{Cov}(Z, Y) = \text{Cov}(AY + \epsilon, Y) = A\Sigma_0$.

Prediction for a single district simply takes the entry corresponding to the district in the distribution $Y \mid Z$, a marginally Normal distribution, and computes the probability that the Normal random variable is greater than 0.5, which is equal to the (marginal) probability that a Republican wins in the district. Prediction for the entire race (as in Figures 7 to 9) is done by drawing $\tilde{y}$ from $Y \mid Z$ and computing the number of entries for which $\tilde{y}_j > 0.5$, as the seats that Republicans win among the competitive seats.

3 Data

3.1 Features for constructing a prior (Section 2)

We first consider a range of potentially relevant features to uncover an overall trend in Midterm Election results since 2010. We collect, process, then use a second-order polynomial basis transform on the features to discover nonlinear correlations in constructing our prior. The following features were used to inform our prior: Republican or Democratic incumbency, gender, voting pattern in the previous presidential election, percentage of minorities, presidential approval rating, candidates party alignment with the incumbent president, log of

[11] Filling missing values with state-averages. This is somewhat an over-estimate of $\mathbb{V}(\epsilon_j)$, since the empirical variance should be unbiased for the total variance instead of the between-variance. We do not take the empirical variance directly since we still would like to weigh the polls with larger sample sizes as more informative than those without. Our variance estimates are likely double-counting the within-poll variances, but this enables us to weigh better polls more.

median income, percentage of district population with a Bachelors degree or above, one-hot encoding of each state, and the ratio of search frequency via Google Trends.

One design choice we made in feature selection is to ignore third-party candidates. Since 2010, there was a single Independent candidate (Jo Ann Emerson) who won a seat at the House, and given that there was no evidence to support a sudden surge in third-party candidates in the 2018 election, we focus on the major parties.

To illustrate the relative importance of these featurees, Figure 1 shows the top 10 features in terms of absolute value of weight coefficients. Empirically, we observe that minority percentage, percentage with Bachelor's degree, and presidential approval rating are features that carry the most weight. Note that *rep_to_tot_oct* and *rep_to_tot_nov* also exhibit strong predictive power, which represent the query frequency on Google Trends for the Republican candidate over Republican and Democratic candidates in the first half of October and second half of October, respectively.

3.2 Google Trends

Google Trends, unlike Google Keyword Planner, shows the *relative* popularity of search queries from a scale of 0 to 100. Relative popularity refers to the ratio of a query's search volumne to the sum of the search volumes of all possible queries. This allows for an easier and more intuitive comparison between terms via search across arbitrary periods of time. Moreover, Google Trends allows for a state-level (but not district-level) granularity in search query trends. We take advantage of this to compare how often a candidate's name was searched in within the state compared to his or her opponent for the first half of October versus the second half of October.

There are two considerations with respect to using Google Trends data. One obvious but important point to note is that we do not discern the positivity or negativity of these queries. Instead, we simply observe the total number of queries. Though this may be a noisy estimate of the true support for a candidate, as was the case for the 2016 Presidential Election, we adhere to the notion that *there is no such thing as bad publicity* (Berger, Sorensen, and Rasmussen, 2010) and The Economist's article that Google Trends "provide real-time information that is not available anywhere else". Second, the usage of Google as a way to collect information in 2018 increased significantly compared to previous years. Furthermore, the use of Google is biased in favor of a younger population, and may represent the interest of young voters disproportionately. With these two caveats in mind, however, we find that the inclusion of Google Trends data augments the information of our prior significantly, compared to a prediction without it.

3.3 Preprocessing

In addition to the standard zero-mean and normalization procedure, we employed several techniques to filter, augment, and otherwise preprocess the collected data in order to produce a sensible outcome by making the most out of the data. In the case of gender, the feature set provided by the class was incomplete. In order to fill the missing values, we use the

gender-guesser library to help classify entries in which we lack the information.[12] In the case of a gender neutral names that were not classified in the training data, we populated them with 0.5.

Another feature with a significant number of missing entries was the percentage of self-identifying Republicans versus Democrats in the district. In order to fill the missing entries with a reasonable approximation, we augmented the features with the 2016 State-level political allegiance data if no data was available for the district for any years, while fitting a forward-filling certain years in question for districts that have the data for some years but not others. For other features with a few missing values, we made weak assumptions about the similarity of districts within a given year and the mean within a year to fill the empty entries.

3.4 Polling Data for Posterior Updates

Once we constructed our preliminary model, we collect and utilize two kinds of polling data to update our prior: the generic ballot poll and a district-wide poll. Both the generic and district polling data were collected from FiveThirtyEight. The first is the generic ballot question, which simply asks a respondent whether she would vote for a Democrat or a Republican. FiveThirtyEight claims to use a slow moving generic ballot average, incorporating a larger number of polls at the cost of recency of polls.

Figure 3 from FiveThirtyEight delineates the estimate by the solid lines while the highlighted badwidth around represents the 90% confidence interval in which the polls fall. This is determined by a set of factors, including days from poll to election, size of poll, disagreement with other polls, lopsidedness of a race, among others.[13] The second is the latest district-level polls, also collected by FiveThirtyEight. As we will later discuss in Section 2.3.2, for both the generic as well as the district poll, we use the a combination of FiveThirtyEight's confidence interval, polls' sample size, and empirical between-poll dispersions to estimate within-poll and between-poll variance.[14] Though we have access to metadata such as the simple average error of the polling firm, grading of the quality of the poll, etc., with access to only the most recent polling results for the 2018 Midterm Election, we opt to make use of the provided data that are hold conjugacy properties while being subject to explicit error quantification.

For the generic congressional ballot, we use all polls fetched from FiveThirtyEight since October 15, 2018. For the district-wide congressional polls, we use the most recent three polls in each district. These choices are rather arbitrary, since we cannot validate our choices given we do not yet observe the outcome variable.

[12] Python `gender-guesser` Library

[13] How The FiveThirtyEight Senate Forecast Model Works

[14] We take the view that the variance of polls decomposes to a within and between part—we can think of a poll result to be a scaled Binomial distribution with additive noise. The variance of the sample-size-scaled Binomial is within-poll variance, which represents the poll's sampling error. The variance of the (theorized) additive noise is to represent a betwee-poll variance that explains the dispersion of poll results unpredicted by the purely Binomial model. See Section 2.3.2 for details.

4 Results

4.1 Regression diagnostics

Before discussing the outcome of our prediction, we first verify our initial assumption that the misspecification in (1) is not severe. We do so by plotting a quantile-quantile plot against a Normal distribution in Figure 5. Indeed, through a quick visual inspection of Figure 5, the empirical quantile and the true Normal quantile form a roughly straight line, suggesting our assumptions hold water. Next, we examine the residuals and the fitted values, a standard regression diagnostic. Figure 6 shows that the residuals are approximately centered around 0, and combined with Figure 5, we observe that their spread is also approximately normal and homoskedastic.

4.2 Final Projection and Baseline Comparisons

Our predictions are in Appendix A.

Graphically, our final prediction for the 2018 Midterm Elections is shown in Figure 7. We plot the distributions of our prior, posterior after updating with the generic ballot, and the final posterior after updating with the district-level polls from Section 2.4 across 5×10^4 samples.

As discussed in Sections 2 and 3, our prior is informed by the collected historical data. Reflecting the ground truth trend of Republican dominance in the House, our prior is centered at a Republican majority with the likelihood of Democrats taking the majority at 17%. With subsequent polling updates to our prior, we observe the posterior shift to a Democratic majority with a slightly higher peak. The dramatic shift from prior to posterior highlights the importance of our two-step procedure, where the data is taken to estimate a prior instead of performing predictions directly with training data.[15]

Alternatively, the high degree of updating might lead one to ask whether the prior estimation is needed in the first place, since the polling data seems to drive our results. To illustrate the informative nature of our data-derived prior, we plot the shift in distribution with a large-variance prior $\mathcal{N}(0.5\mathbf{1}, 10I)$ in Figure 9. We see that having a prior informed from data (1) helps our model tremendously with districts for which we do not have polling data and (2) crucially estimates Σ_0, which is important for updating.

As a robustness check, we also plot the updating for both a diagonal specification of Σ_0 and a fully-general specification. Our final posterior presented in Figure 7 makes use of Σ_0 for our prior covariance matrix where the off-diagonal entries are estimated via the empirical variation over time assuming unchanged district boundaries. If we assume prior independence across each district, we have a diagonal Σ_0, which produces the result of Figure 8. We see that having a nondiagonal prior dramatically reduces the certainty in number of Republican seats, compared to the independent prior model in Figure 8. With nondiagonal priors, the update of district-wide polls actually shifts the distribution to the right, an effect similar to that discussed in Figure 9.

[15] The probabilities displayed here might change on a daily basis given new polling data. See Appendix A for our final estimates.

4.3 Comparing against the state-of-the-art

Figure 10 provides FiveThirtyEight's predictions on the day before the Election as a baseline against which we can compare our own predictions. We also map our prediction on the district level in Figure 11 and compare our results against that of FiveThirtyEight in Figure 12. Note the wider tails in FiveThirtyEight's prediction and a more favorable mean for the Democrats than our predictions. FiveThirtyEight provides much more uncertainty than our estimates, which may come from a better model of conditional correlation across polls, better understanding of polling variance, and better data for prediction of "fundamentals." Directionally, our projections for each race are quite similar to that of FiveThirtyEight's, and our median Democratic seats is also similar to that of FiveThirtyEight.

5 Conclusion, Discussion, and Future Considerations

Our results, in Appendix A, are broadly inline with those of professional forecasters, except for the estimated variance of our results. Our model is simple to estimate and extend, requiring little computational power in parts other than the elastic net in (Estimation for β_0), and can easily incorporate more data, features, or polls. Most importantly, as demonstrated by Figures 7 and 9, our Bayesian second-step estimates effectively take into account real-time polling information, while our empirically estimated first-step prior fills in places where the polls are lacking.

Our most important shortcoming—if we believe that professional forecasters have better models than ours—is that the variance we obtain is too small. The RMSE of the elastic net in Equation (Estimation for β_0) is about 7%. Thus the diagonal for Σ_0 is centered around $(7\%)^2$. This figure is reduced to about $(6\%)^2$ after the generic-ballot update and to $(5\%)^2$ after both rounds of Bayesian updates, meaning that a two-standard-deviation interval is $\pm 10\%$ in Republican vote-share for the average race, which appears, a priori, to be quite generous and conservative; nonetheless, the variance produced by FiveThirtyEight's forecast is probably a better measure. There are a few ways to adjust the model to be more realistic and increase variance. The prior variance estimates in Σ_0 does not consider the variance introduced by possible generalization errors—if we believe in some time-series model of political leanings, then there are necessarily some between-year variance that we failed to account here. A second consideration for variance could be the (probably false) assumption that polls are independently distributed conditional on the true proportion Y. We know from anecdotal evidence that polls are quite correlated, as was the conventional wisdom after Trump's 2016 election shock. At the minimum, we should expect that polls by the same pollsters are correlated, and polls by pollsters with political agendas are even biased.[16] A more sophisticated updating model, which takes into account the off-diagonal terms in Σ_Z, would probably increase variance of the posterior, as the information provided by the polls is no longer as strong as it is in a world where polls are assumed (conditionally) independent.

Moreover, there are a number of ways through which our prediction in the first-stage

[16] FiveThirtyEight provides some measures of poll quality, which we ignore in our analysis, but they might be important for the future.

could have been more informed. On the data level, we could have integrated a more all-encompassing Google Trends search. Instead of simply gauging interest for candidates' names, a more holistic understanding can be had by searching Trends results for political topics (i.e. abortion, health care, immigration, etc.) and adding those keywords as features.

Another source of improvement could be in model construction and comparison. Instead of using a regularized regression, we could have used smoothing splines of the form

$$\widehat{f} = \arg\min_f \mathrm{RSS}(f, \lambda) = \arg\min_f \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt$$

in order to impose a smoothness to the overarching prediction function without explicitly using a second degree polynomial basis transform. This has the property of maintaining interpretability while also restricting the degrees of freedom. The other extreme would be to make use of the neural networks model, although we opted to trade the advantages of flexibility in non-linear transformations with interpretability and fewer parameters, many of which could materially change the outcome of the performance of neural nets (e.g. starting values, optimization function, input scaling, among others). A valuable extension could be to compare the performance of the aforementioned models and to use ensemble learning to obtain a better predictive performance that incorporates various approaches.

References

Berger, Jonah, Alan T Sorensen, and Scott J Rasmussen. 2010. "Positive effects of negative publicity: When negative reviews increase sales." *Marketing Science* 29 (5):815–827.

Grün, Bettina, Ioannis Kosmidis, and Achim Zeileis. 2011. "Extended beta regression in R: shaken, stirred, mixed, and partitioned." Tech. rep., Working Papers in Economics and Statistics.

Kyung, Minjung, Jeff Gill, Malay Ghosh, George Casella et al. 2010. "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis* 5 (2):369–411.

Zou, Hui, Trevor Hastie, Robert Tibshirani et al. 2007. "On the degrees of freedom of the lasso." *The Annals of Statistics* 35 (5):2173–2192.

A District-by-district projections

The probability that Democrats take House majority for model with prior is 0.3409. The probability that Democrats take House majority for model with posterior after generic ballot is 1.0000. The probability that Democrats take House majority for model with posterior is 0.8564.

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Alabama-01	Byrne (R)	1.0000	0.9997	0.0003
Alabama-02	Roby (R)	0.9883	0.9780	0.0220
Alabama-03	Rogers (R)	1.0000	0.9983	0.0017
Alabama-04	Aderholt (R)	1.0000	1.0000	0.0000
Alabama-05	Brooks (R)	1.0000	0.9991	0.0009
Alabama-06	Palmer (R)	1.0000	0.9999	0.0001
Alabama-07	Sewell (D)	1.0000	—	1.0000
Alaska-00	Young (R)	1.0000	0.6538	0.3462
Arizona-01	O'Halleran (D)	1.0000	0.1088	0.8912
Arizona-02	Kirkpatrick (D)	0.9520	0.0355	0.9645
Arizona-03	Grijalva (D)	1.0000	0.0001	0.9999
Arizona-04	Gosar (R)	1.0000	0.9994	0.0006
Arizona-05	Biggs (R)	1.0000	0.9991	0.0009
Arizona-06	Schweikert (R)	1.0000	0.9081	0.0919
Arizona-07	Gallego (D)	1.0000	—	1.0000
Arizona-08	Lesko (R)	1.0000	0.7767	0.2233
Arizona-09	Stanton (D)	0.8183	0.0051	0.9949
Arkansas-01	Crawford (R)	1.0000	0.9995	0.0005
Arkansas-02	Hill (R)	0.9995	0.8622	0.1378
Arkansas-03	Womack (R)	1.0000	0.9991	0.0009
Arkansas-04	Westerman (R)	1.0000	0.9997	0.0003
California-01	LaMalfa (R)	1.0000	0.7889	0.2111
California-02	Huffman (D)	1.0000	0.0000	1.0000
California-03	Garamendi (D)	0.9997	0.0003	0.9997
California-04	McClintock (R)	0.8708	0.8476	0.1524
California-05	Thompson (D)	1.0000	—	0.9998
California-06	Jefferson (D)	1.0000	—	0.0000
California-07	Bera (D)	1.0000	0.0049	0.9951
California-08	Donnelly (R)	1.0000	0.0343	—
California-09	McNerney (D)	1.0000	0.0007	0.9993
California-10	Denham (R)	0.8140	0.2296	0.7704
California-11	DeSaulnier (D)	1.0000	0.0000	1.0000
California-12	Pelosi (D)	1.0000	0.0000	1.0000
California-13	Lee (D)	1.0000	—	1.0000
California-14	Speier (D)	1.0000	0.0000	1.0000

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
California-15	Swalwell (D)	1.0000	0.0000	1.0000
California-16	Costa (D)	1.0000	0.0169	0.9831
California-17	Khanna (D)	1.0000	0.0000	1.0000
California-18	Eshoo (D)	1.0000	0.0000	1.0000
California-19	Lofgren (D)	1.0000	0.0000	1.0000
California-20	Panetta (D)	1.0000	—	0.9997
California-21	Valadao (R)	0.9993	0.7948	0.2052
California-22	Nunes (R)	0.9998	0.9555	0.0445
California-23	McCarthy (R)	0.9965	0.9994	0.0006
California-24	Carbajal (D)	0.9964	0.0146	0.9854
California-25	Hill (D)	0.7649	0.3692	0.6308
California-26	Brownley (D)	0.9998	0.0002	0.9998
California-27	Witt (D)	1.0000	—	0.0000
California-28	Schiff (D)	1.0000	0.0000	1.0000
California-29	Crdenas (D)	0.9996	0.0000	1.0000
California-30	Sherman (D)	1.0000	0.0000	1.0000
California-31	Aguilar (D)	0.9998	0.0012	0.9988
California-32	Napolitano (D)	1.0000	0.0000	1.0000
California-33	Lieu (D)	1.0000	0.0000	1.0000
California-34	Gomez (D)	1.0000	—	0.9999
California-35	Torres (D)	1.0000	0.0000	1.0000
California-36	Ruiz (D)	1.0000	0.0015	0.9985
California-37	Bass (D)	1.0000	0.0000	1.0000
California-38	Snchez (D)	1.0000	0.0000	1.0000
California-39	Cisneros (D)	1.0000	0.4202	0.5798
California-40	Roybal-Allard (D)	1.0000	—	0.9997
California-41	Takano (D)	1.0000	0.0001	0.9999
California-42	Peacock (D)	0.8153	0.9757	0.0243
California-43	Waters (D)	1.0000	0.0000	1.0000
California-44	Brown (D)	1.0000	—	0.0000
California-45	Porter (D)	0.8664	0.3768	0.6232
California-46	Correa (D)	1.0000	0.0000	1.0000
California-47	Lowenthal (D)	1.0000	0.0001	0.9999
California-48	Rouda (D)	0.6622	0.4365	0.5635
California-49	Levin (D)	0.5186	0.0401	0.9599
California-50	Hunter (R)	0.9848	0.7822	0.2178
California-51	Vargas (D)	1.0000	0.0000	1.0000
California-52	Peters (D)	1.0000	0.0004	0.9996
California-53	Davis (D)	1.0000	0.0000	1.0000
Colorado-01	DeGette (D)	1.0000	0.0000	1.0000
Colorado-02	Yu (R)	0.8490	0.0003	0.9997

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Colorado-03	Tipton (R)	0.9854	0.7980	0.2020
Colorado-04	Buck (R)	0.9979	0.9584	0.0416
Colorado-05	Lamborn (R)	0.8826	0.9791	0.0209
Colorado-06	Crow (D)	0.9928	0.1175	0.8825
Colorado-07	Perlmutter (D)	0.9995	0.0006	0.9994
Connecticut-01	Larson (D)	1.0000	0.0001	0.9999
Connecticut-02	Courtney (D)	0.9999	0.0003	0.9997
Connecticut-03	DeLauro (D)	1.0000	0.0000	1.0000
Connecticut-04	Himes (D)	1.0000	0.0005	0.9995
Connecticut-05	Hayes (D)	0.9995	0.0255	0.9745
Delaware-00	Rochester (D)	1.0000	0.0019	0.9981
Florida-01	Gaetz (R)	0.9976	1.0000	0.0000
Florida-02	Dunn (R)	0.9953	0.9997	0.0003
Florida-03	Yoho (R)	0.9057	0.9858	0.0142
Florida-04	Rutherford (R)	1.0000	0.9999	0.0001
Florida-05	Lawson (D)	1.0000	0.0000	1.0000
Florida-06	Waltz (R)	0.9657	0.7376	0.2624
Florida-07	Murphy (D)	1.0000	0.0532	0.9468
Florida-08	Posey (R)	0.9050	0.9934	0.0066
Florida-09	Soto (D)	1.0000	0.0154	0.9846
Florida-10	Demings (D)	1.0000	—	1.0000
Florida-11	Webster (R)	0.9993	0.9997	0.0003
Florida-12	Bilirakis (R)	0.9997	0.9949	0.0051
Florida-13	Crist (D)	1.0000	0.0013	0.9987
Florida-14	Castor (D)	1.0000	—	1.0000
Florida-15	Carlson (D)	0.5556	0.5823	0.4177
Florida-16	Buchanan (R)	0.9991	0.8592	0.1408
Florida-17	Steube (R)	0.9565	0.9984	0.0016
Florida-18	Mast (R)	0.9963	0.9099	0.0901
Florida-19	Rooney (R)	1.0000	0.9908	0.0092
Florida-20	Hastings (D)	1.0000	—	1.0000
Florida-21	Frankel (D)	1.0000	—	1.0000
Florida-22	Deutch (D)	1.0000	0.0004	0.9996
Florida-23	Schultz (D)	1.0000	0.0003	0.9997
Florida-24	Wilson (D)	1.0000	—	1.0000
Florida-25	Diaz-Balart (R)	0.9998	0.7285	0.2715
Florida-26	Curbelo (R)	0.6236	0.4418	0.5582
Florida-27	Salazar (R)	0.6744	0.1519	0.8481
Georgia-01	Carter (R)	1.0000	0.9986	0.0014
Georgia-02	Bishop (D)	1.0000	0.0001	0.9999
Georgia-03	Ferguson (R)	1.0000	1.0000	0.0000

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Georgia-04	Johnson (D)	1.0000	0.0000	1.0000
Georgia-05	Lewis (D)	1.0000	—	1.0000
Georgia-06	Handel (R)	0.9893	0.5068	0.4932
Georgia-07	Woodall (R)	0.8397	0.8412	0.1588
Georgia-08	Scott (R)	1.0000	1.0000	—
Georgia-09	Collins (R)	1.0000	1.0000	0.0000
Georgia-10	Hice (R)	1.0000	0.9999	0.0001
Georgia-11	Loudermilk (R)	1.0000	0.9998	0.0002
Georgia-12	Allen (R)	1.0000	0.9883	0.0117
Georgia-13	Scott (D)	1.0000	0.0000	1.0000
Georgia-14	Graves (R)	1.0000	1.0000	0.0000
Hawaii-01	Case (D)	1.0000	0.0000	1.0000
Hawaii-02	Gabbard (D)	1.0000	0.0000	1.0000
Idaho-01	Fulcher (R)	1.0000	0.9944	0.0056
Idaho-02	Simpson (R)	1.0000	0.9991	0.0009
Illinois-01	Rush (D)	1.0000	0.0000	1.0000
Illinois-02	Kelly (D)	1.0000	0.0000	1.0000
Illinois-03	Lipinski (D)	0.9906	0.0001	0.9999
Illinois-04	Garcia (D)	0.9999	0.0000	1.0000
Illinois-05	Quigley (D)	1.0000	0.0000	1.0000
Illinois-06	Casten (D)	0.8652	0.5226	0.4774
Illinois-07	Davis (D)	1.0000	0.0000	1.0000
Illinois-08	Krishnamoorthi (D)	1.0000	0.0001	0.9999
Illinois-09	Schakowsky (D)	1.0000	0.0000	1.0000
Illinois-10	Schneider (D)	1.0000	0.0002	0.9998
Illinois-11	Foster (D)	1.0000	0.0001	0.9999
Illinois-12	Bost (R)	0.9868	0.7268	0.2732
Illinois-13	Davis (R)	0.8623	0.7243	0.2757
Illinois-14	Underwood (D)	0.5365	0.3115	0.6885
Illinois-15	Shimkus (R)	1.0000	1.0000	0.0000
Illinois-16	Kinzinger (R)	0.9992	0.9912	0.0088
Illinois-17	Bustos (D)	1.0000	0.0004	0.9996
Illinois-18	LaHood (R)	0.9999	0.9999	0.0001
Indiana-01	Visclosky (D)	0.9986	0.0000	1.0000
Indiana-02	Walorski (R)	1.0000	0.9604	0.0396
Indiana-03	Banks (R)	1.0000	0.9977	0.0023
Indiana-04	Baird (R)	0.9945	0.9984	0.0016
Indiana-05	Brooks (R)	0.9997	0.9623	0.0377
Indiana-06	Pence (R)	1.0000	0.9996	0.0004
Indiana-07	Carson (D)	1.0000	0.0000	1.0000
Indiana-08	Bucshon (R)	1.0000	0.9970	0.0030

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Indiana-09	Hollingsworth (R)	1.0000	0.8336	0.1664
Iowa-01	Blum (R)	0.9880	0.0440	0.9560
Iowa-02	Loebsack (D)	1.0000	0.0107	0.9893
Iowa-03	Axne (D)	0.7398	0.3039	0.6961
Iowa-04	King (R)	1.0000	0.8521	0.1479
Kansas-01	Marshall (R)	1.0000	0.9986	0.0014
Kansas-02	Watkins (R)	0.9994	0.3795	0.6205
Kansas-03	Davids (D)	0.8016	0.1524	0.8476
Kansas-04	Estes (R)	0.9999	0.9932	0.0068
Kentucky-01	Comer (R)	1.0000	0.9999	0.0001
Kentucky-02	Guthrie (R)	1.0000	0.9993	0.0007
Kentucky-03	Yarmuth (D)	1.0000	0.0015	0.9985
Kentucky-04	Massie (R)	1.0000	0.9997	0.0003
Kentucky-05	Rogers (R)	1.0000	1.0000	0.0000
Kentucky-06	Barr (R)	1.0000	0.5450	0.4550
Louisiana-01	Scalise (R)	1.0000	1.0000	0.0000
Louisiana-02	Richmond (D)	1.0000	—	1.0000
Louisiana-03	Guillory (R)	1.0000	0.0033	0.0000
Louisiana-04	Johnson (R)	0.9999	0.9994	0.0005
Louisiana-05	Abraham (R)	1.0000	0.9998	0.0002
Louisiana-06	Graves (R)	1.0000	0.9999	0.0000
Maine-01	Pingree (D)	1.0000	0.0003	0.9997
Maine-02	Poliquin (R)	0.8888	0.3701	0.6299
Maryland-01	Harris (R)	0.9998	0.9905	0.0095
Maryland-02	Ruppersberger (D)	1.0000	0.0001	0.9999
Maryland-03	Sarbanes (D)	1.0000	0.0001	0.9999
Maryland-04	Brown (D)	1.0000	0.0000	1.0000
Maryland-05	Hoyer (D)	1.0000	0.0000	1.0000
Maryland-06	Trone (D)	0.9999	0.0051	0.9949
Maryland-07	Cummings (D)	1.0000	0.0000	1.0000
Maryland-08	Raskin (D)	1.0000	0.0000	1.0000
Massachusetts-01	Neal (D)	1.0000	—	1.0000
Massachusetts-02	McGovern (D)	1.0000	0.0000	1.0000
Massachusetts-03	Trahan (D)	1.0000	0.0004	0.9996
Massachusetts-04	Kennedy (D)	1.0000	—	1.0000
Massachusetts-05	Clark (D)	1.0000	0.0000	1.0000
Massachusetts-06	Moulton (D)	1.0000	0.0001	0.9999
Massachusetts-07	Pressley (D)	1.0000	—	1.0000
Massachusetts-08	Lynch (D)	1.0000	—	1.0000
Massachusetts-09	Keating (D)	1.0000	0.0027	0.9973
Michigan-01	Bergman (R)	1.0000	0.9257	0.0743

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Michigan-02	Huizenga (R)	0.9916	0.9398	0.0602
Michigan-03	Amash (R)	0.9938	0.9815	0.0185
Michigan-04	Moolenaar (R)	1.0000	0.9986	0.0014
Michigan-05	Kildee (D)	0.9999	0.0003	0.9997
Michigan-06	Upton (R)	0.9287	0.7738	0.2262
Michigan-07	Walberg (R)	0.9997	0.5875	0.4125
Michigan-08	Bishop (R)	0.7904	0.3284	0.6716
Michigan-09	Levin (D)	0.8998	0.0034	0.9966
Michigan-10	Mitchell (R)	1.0000	0.9988	0.0012
Michigan-11	Stevens (D)	0.8728	0.2063	0.7937
Michigan-12	Dingell (D)	1.0000	0.0000	1.0000
Michigan-13	Tlaib (D)	1.0000	—	1.0000
Michigan-14	Lawrence (D)	1.0000	0.0000	1.0000
Minnesota-01	Hagedorn (R)	0.9984	0.4532	0.5468
Minnesota-02	Lewis (R)	0.8386	0.1361	0.8639
Minnesota-03	Phillips (D)	0.9963	0.1549	0.8451
Minnesota-04	McCollum (D)	1.0000	0.0000	1.0000
Minnesota-05	Omar (D)	1.0000	0.0000	1.0000
Minnesota-06	Emmer (R)	1.0000	0.9989	0.0011
Minnesota-07	Peterson (D)	0.9964	0.0670	0.9330
Minnesota-08	Stauber (R)	0.9467	0.8113	0.1887
Mississippi-01	Kelly (R)	1.0000	0.9995	0.0005
Mississippi-02	Thompson (D)	1.0000	—	1.0000
Mississippi-03	Guest (R)	1.0000	0.9968	0.0032
Mississippi-04	Palazzo (R)	1.0000	0.9960	0.0040
Missouri-01	Clay (D)	1.0000	0.0000	1.0000
Missouri-02	Wagner (R)	1.0000	0.8422	0.1578
Missouri-03	Luetkemeyer (R)	1.0000	0.9999	0.0001
Missouri-04	Hartzler (R)	1.0000	0.9994	0.0006
Missouri-05	Cleaver (D)	1.0000	0.0005	0.9995
Missouri-06	Graves (R)	1.0000	0.9997	0.0003
Missouri-07	Long (R)	1.0000	0.9999	0.0001
Missouri-08	Smith (R)	1.0000	1.0000	0.0000
Montana-00	Gianforte (R)	1.0000	0.7581	0.2419
Nebraska-01	Fortenberry (R)	1.0000	0.9977	0.0023
Nebraska-02	Bacon (R)	1.0000	0.5834	0.4166
Nebraska-03	Smith (R)	1.0000	1.0000	0.0000
Nevada-01	Titus (D)	1.0000	0.0001	0.9999
Nevada-02	Amodei (R)	0.9998	0.9977	0.0023
Nevada-03	Lee (D)	0.9912	0.1194	0.8806
Nevada-04	Horsford (D)	0.9635	0.1230	0.8770

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
New Hampshire-01	Edwards (R)	1.0000	0.1087	0.8913
New Hampshire-02	Kuster (D)	0.8995	0.0035	0.9965
New Jersey-01	Norcross (D)	1.0000	0.0000	1.0000
New Jersey-02	Van Drew (D)	0.9998	0.0208	0.9792
New Jersey-03	Kim (D)	0.6173	0.4538	0.5462
New Jersey-04	Smith (R)	0.9997	0.9294	0.0706
New Jersey-05	Gottheimer (D)	1.0000	0.0133	0.9867
New Jersey-06	Pallone (D)	1.0000	0.0001	0.9999
New Jersey-07	Malinowski (D)	0.5280	0.2211	0.7789
New Jersey-08	Sires (D)	1.0000	0.0000	1.0000
New Jersey-09	Pascrell (D)	1.0000	0.0000	1.0000
New Jersey-10	Payne (D)	1.0000	0.0000	1.0000
New Jersey-11	Sherrill (D)	0.9990	0.1483	0.8517
New Jersey-12	Coleman (D)	1.0000	0.0000	1.0000
New Mexico-01	Haaland (D)	1.0000	0.0213	0.9787
New Mexico-02	Herrell (R)	0.7309	0.5579	0.4421
New Mexico-03	Lujan (D)	1.0000	0.0005	0.9995
New York-01	Zeldin (R)	1.0000	0.9346	0.0654
New York-02	King (R)	0.5503	0.7293	0.2707
New York-03	Suozzi (D)	1.0000	0.0032	0.9968
New York-04	Rice (D)	1.0000	0.0004	0.9996
New York-05	Meeks (D)	1.0000	—	1.0000
New York-06	Meng (D)	1.0000	—	0.9999
New York-07	Velzquez (D)	1.0000	—	1.0000
New York-08	Jeffries (D)	1.0000	—	1.0000
New York-09	Clarke (D)	1.0000	0.0000	1.0000
New York-10	Nadler (D)	1.0000	0.0000	1.0000
New York-11	Donovan (R)	0.7638	0.7644	0.2356
New York-12	Maloney (D)	1.0000	0.0000	1.0000
New York-13	Espaillat (D)	1.0000	0.0000	1.0000
New York-14	Ocasio-Cortez (D)	1.0000	0.0000	1.0000
New York-15	Serrano (D)	1.0000	0.0000	1.0000
New York-16	Engel (D)	1.0000	—	1.0000
New York-17	Lowey (D)	1.0000	—	0.9999
New York-18	Maloney (D)	1.0000	0.0076	0.9924
New York-19	Faso (R)	0.9469	0.3900	0.6100
New York-20	Tonko (D)	1.0000	0.0000	1.0000
New York-21	Stefanik (R)	1.0000	0.8936	0.1064
New York-22	Tenney (R)	1.0000	0.4800	0.5200
New York-23	Reed (R)	0.9887	0.8302	0.1698
New York-24	Katko (R)	0.9157	0.8211	0.1789

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
New York-25	Morelle (D)	0.9917	0.0100	0.9900
New York-26	Higgins (D)	1.0000	0.0000	1.0000
New York-27	Collins (R)	1.0000	0.7583	0.2417
North Carolina-01	Butterfield (D)	1.0000	0.0000	1.0000
North Carolina-02	Holding (R)	0.9123	0.8479	0.1521
North Carolina-03	Jones (R)	1.0000	1.0000	—
North Carolina-04	Price (D)	1.0000	0.0000	1.0000
North Carolina-05	Foxx (R)	1.0000	0.9602	0.0398
North Carolina-06	Walker (R)	0.6335	0.8706	0.1294
North Carolina-07	Rouzer (R)	0.7828	0.8177	0.1823
North Carolina-08	Hudson (R)	0.9633	0.8771	0.1229
North Carolina-09	McCready (D)	0.9749	0.5302	0.4698
North Carolina-10	McHenry (R)	0.9988	0.9983	0.0017
North Carolina-11	Meadows (R)	1.0000	0.9973	0.0027
North Carolina-12	Adams (D)	1.0000	0.0000	1.0000
North Carolina-13	Budd (R)	0.9943	0.6132	0.3868
North Dakota-00	Armstrong (R)	1.0000	0.9989	0.0011
Ohio-01	Chabot (R)	0.7403	0.7996	0.2004
Ohio-02	Wenstrup (R)	1.0000	0.9757	0.0243
Ohio-03	Beatty (D)	1.0000	0.0000	1.0000
Ohio-04	Jordan (R)	0.9999	0.9906	0.0094
Ohio-05	Latta (R)	0.9988	0.9985	0.0015
Ohio-06	Johnson (R)	1.0000	0.9999	0.0001
Ohio-07	Gibbs (R)	1.0000	0.9634	0.0366
Ohio-08	Davidson (R)	0.9999	0.9997	0.0003
Ohio-09	Kaptur (D)	1.0000	0.0000	1.0000
Ohio-10	Turner (R)	0.7941	0.9019	0.0981
Ohio-11	Fudge (D)	1.0000	0.0000	1.0000
Ohio-12	Balderson (R)	0.9840	0.6549	0.3450
Ohio-13	Ryan (D)	1.0000	0.0001	0.9999
Ohio-14	Joyce (R)	0.9954	0.8463	0.1537
Ohio-15	Stivers (R)	0.9996	0.9791	0.0209
Ohio-16	Gonzalez (R)	0.9971	0.9169	0.0831
Oklahoma-01	Hern (R)	1.0000	0.9984	0.0016
Oklahoma-02	Mullin (R)	1.0000	0.9981	0.0019
Oklahoma-03	Lucas (R)	1.0000	1.0000	0.0000
Oklahoma-04	Cole (R)	1.0000	1.0000	0.0000
Oklahoma-05	Russell (R)	1.0000	0.8636	0.1364
Oregon-01	Bonamici (D)	1.0000	0.0001	0.9999
Oregon-02	Walden (R)	1.0000	0.9901	0.0099
Oregon-03	Blumenauer (D)	1.0000	0.0000	1.0000

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Oregon-04	DeFazio (D)	1.0000	0.0047	0.9953
Oregon-05	Schrader (D)	1.0000	0.1449	0.8551
Pennsylvania-01	Wallace (D)	1.0000	0.5736	0.4264
Pennsylvania-02	Boyle (D)	1.0000	0.0000	1.0000
Pennsylvania-03	Evans (D)	0.9289	0.0000	1.0000
Pennsylvania-04	David (R)	0.9978	0.0010	0.9990
Pennsylvania-05	Kim (R)	0.8269	0.0001	0.9999
Pennsylvania-06	Houlahan (D)	0.6856	0.0083	0.9917
Pennsylvania-07	Nothstein (R)	0.7259	0.0490	0.9510
Pennsylvania-08	Cartwright (D)	1.0000	0.0372	0.9628
Pennsylvania-09	Meuser (R)	1.0000	0.9982	0.0018
Pennsylvania-10	Perry (R)	0.9998	0.6782	0.3218
Pennsylvania-11	Smucker (R)	0.9648	0.9327	0.0673
Pennsylvania-12	Marino (R)	1.0000	0.9997	0.0003
Pennsylvania-13	Ottaway (D)	1.0000	0.9996	0.0004
Pennsylvania-14	Boerio (D)	1.0000	0.9930	0.0070
Pennsylvania-15	Thompson (R)	0.9978	1.0000	0.0000
Pennsylvania-16	Kelly (R)	0.9999	0.8753	0.1247
Pennsylvania-17	Lamb (D)	0.8497	0.0428	0.9572
Pennsylvania-18	Doyle (D)	1.0000	—	1.0000
Rhode Island-01	Cicilline (D)	1.0000	0.0000	1.0000
Rhode Island-02	Langevin (D)	1.0000	0.0001	0.9999
South Carolina-01	Arrington (R)	0.9795	0.9129	0.0871
South Carolina-02	Wilson (R)	1.0000	0.9935	0.0065
South Carolina-03	Duncan (R)	1.0000	0.9996	0.0004
South Carolina-04	Timmons (R)	0.9998	0.9999	0.0001
South Carolina-05	Norman (R)	0.9990	0.9874	0.0125
South Carolina-06	Clyburn (D)	1.0000	0.0000	1.0000
South Carolina-07	Rice (R)	0.9998	0.9937	0.0063
South Dakota-00	Johnson (R)	0.9997	0.9921	0.0079
Tennessee-01	Roe (R)	1.0000	1.0000	0.0000
Tennessee-02	Burchett (R)	1.0000	0.9998	0.0002
Tennessee-03	Fleischmann (R)	1.0000	0.9996	0.0004
Tennessee-04	DesJarlais (R)	1.0000	0.9941	0.0058
Tennessee-05	Cooper (D)	1.0000	0.0001	0.9999
Tennessee-06	Rose (R)	1.0000	1.0000	0.0000
Tennessee-07	Green (R)	1.0000	0.9999	0.0001
Tennessee-08	Kustoff (R)	1.0000	0.9999	0.0001
Tennessee-09	Cohen (D)	1.0000	0.0000	1.0000
Texas-01	Gohmert (R)	1.0000	1.0000	0.0000
Texas-02	Litton (D)	0.8066	0.8955	0.1045

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Texas-03	Burch (D)	0.9857	0.9927	0.0073
Texas-04	Ratcliffe (R)	1.0000	1.0000	0.0000
Texas-05	Wood (D)	0.9261	0.9996	0.0004
Texas-06	Sanchez (D)	0.8416	0.9436	0.0564
Texas-07	Culberson (R)	0.6269	0.4862	0.5138
Texas-08	Brady (R)	1.0000	1.0000	0.0000
Texas-09	Green (D)	1.0000	—	1.0000
Texas-10	McCaul (R)	0.9791	0.9381	0.0619
Texas-11	Conaway (R)	1.0000	1.0000	0.0000
Texas-12	Granger (R)	1.0000	1.0000	0.0000
Texas-13	Thornberry (R)	1.0000	1.0000	0.0000
Texas-14	Weber (R)	1.0000	0.9921	0.0079
Texas-15	Gonzalez (D)	1.0000	0.0003	0.9997
Texas-16	Escobar (D)	0.5255	0.0000	1.0000
Texas-17	Flores (R)	0.9998	0.9979	0.0021
Texas-18	Lee (D)	1.0000	0.0000	1.0000
Texas-19	Arrington (R)	0.9996	1.0000	0.0000
Texas-20	Castro (D)	1.0000	—	0.9998
Texas-21	Roy (R)	0.9784	0.8519	0.1481
Texas-22	Olson (R)	0.7838	0.8043	0.1957
Texas-23	Hurd (R)	0.9945	0.7779	0.2221
Texas-24	Marchant (R)	0.6252	0.9671	0.0329
Texas-25	Williams (R)	1.0000	0.8856	0.1144
Texas-26	Burgess (R)	0.9999	0.9994	0.0006
Texas-27	Cloud (R)	1.0000	0.9986	0.0014
Texas-28	Cuellar (D)	1.0000	—	0.9998
Texas-29	Garcia (D)	0.8336	0.0000	1.0000
Texas-30	Johnson (D)	1.0000	—	1.0000
Texas-31	Carter (R)	1.0000	0.9329	0.0671
Texas-32	Allred (D)	0.6738	0.6316	0.3684
Texas-33	Veasey (D)	1.0000	0.0000	1.0000
Texas-34	Vela (D)	0.9332	0.0003	0.9997
Texas-35	Doggett (D)	1.0000	0.0000	1.0000
Texas-36	Babin (R)	1.0000	1.0000	0.0000
Utah-01	Bishop (R)	1.0000	0.9999	0.0001
Utah-02	Stewart (R)	1.0000	0.9837	0.0162
Utah-03	Curtis (R)	1.0000	0.9998	0.0002
Utah-04	Love (R)	1.0000	0.3857	0.6143
Vermont-00	Welch (D)	1.0000	0.0000	1.0000
Virginia-01	Wittman (R)	0.8042	0.9719	0.0281
Virginia-02	Taylor (R)	0.8683	0.6713	0.3287

Continued on next page

District	Projected Winner	Probability	$\mathbb{P}_{538}(R)$	$\mathbb{P}_{538}(D)$
Virginia-03	Scott (D)	1.0000	—	1.0000
Virginia-04	McEachin (D)	1.0000	0.0006	0.9994
Virginia-05	Riggleman (R)	0.7395	0.5438	0.4562
Virginia-06	Cline (R)	1.0000	0.9947	0.0053
Virginia-07	Brat (R)	0.5659	0.5642	0.4358
Virginia-08	Beyer (D)	1.0000	0.0000	1.0000
Virginia-09	Griffith (R)	1.0000	0.9907	0.0093
Virginia-10	Wexton (D)	0.8488	0.1057	0.8943
Virginia-11	Connolly (D)	1.0000	0.0000	1.0000
Washington-01	DelBene (D)	1.0000	0.0006	0.9994
Washington-02	Larsen (D)	1.0000	—	0.9998
Washington-03	Beutler (R)	1.0000	0.7464	0.2536
Washington-04	Newhouse (R)	0.8927	0.9963	0.0037
Washington-05	Rodgers (R)	0.9985	0.7615	0.2385
Washington-06	Kilmer (D)	1.0000	0.0002	0.9998
Washington-07	Jayapal (D)	1.0000	0.0000	1.0000
Washington-08	Rossi (R)	0.9682	0.3281	0.6719
Washington-09	Smith (D)	1.0000	—	0.0147
Washington-10	Heck (D)	1.0000	0.0003	0.9997
West Virginia-01	McKinley (R)	1.0000	0.9996	0.0004
West Virginia-02	Mooney (R)	1.0000	0.9370	0.0630
West Virginia-03	Miller (R)	1.0000	0.9268	0.0732
Wisconsin-01	Steil (R)	0.5794	0.7825	0.2175
Wisconsin-02	Pocan (D)	1.0000	—	1.0000
Wisconsin-03	Kind (D)	1.0000	0.0026	0.9974
Wisconsin-04	Moore (D)	1.0000	0.0000	1.0000
Wisconsin-05	Sensenbrenner (R)	0.8776	0.9899	0.0101
Wisconsin-06	Grothman (R)	0.9544	0.9528	0.0472
Wisconsin-07	Duffy (R)	0.9999	0.9937	0.0063
Wisconsin-08	Gallagher (R)	0.9978	0.9960	0.0040
Wyoming-00	Cheney (R)	1.0000	0.9998	0.0002

B Figures

minority_percentage_x_Maine	-0.547498
percent_bachelor_or_above_x_Mississippi	0.365465
minority_percentage	-0.355149
percent_bachelor_or_above_x_Alabama	0.300776
rep_to_tot_oct_x_Hawaii	0.278641
percent_bachelor_or_above_x_percent_bachelor_or_above	-0.266805
pres_approval_x_Oklahoma	0.187379
rep_to_tot_nov_x_Vermont	-0.182223
percent_bachelor_or_above_x_Massachusetts	-0.167371
minority_percentage_x_minority_percentage	-0.155483

Figure 1: Some feature coefficients. Positive coefficients roughly mean "associated with Republicans winning," and negative coefficients for Democrats

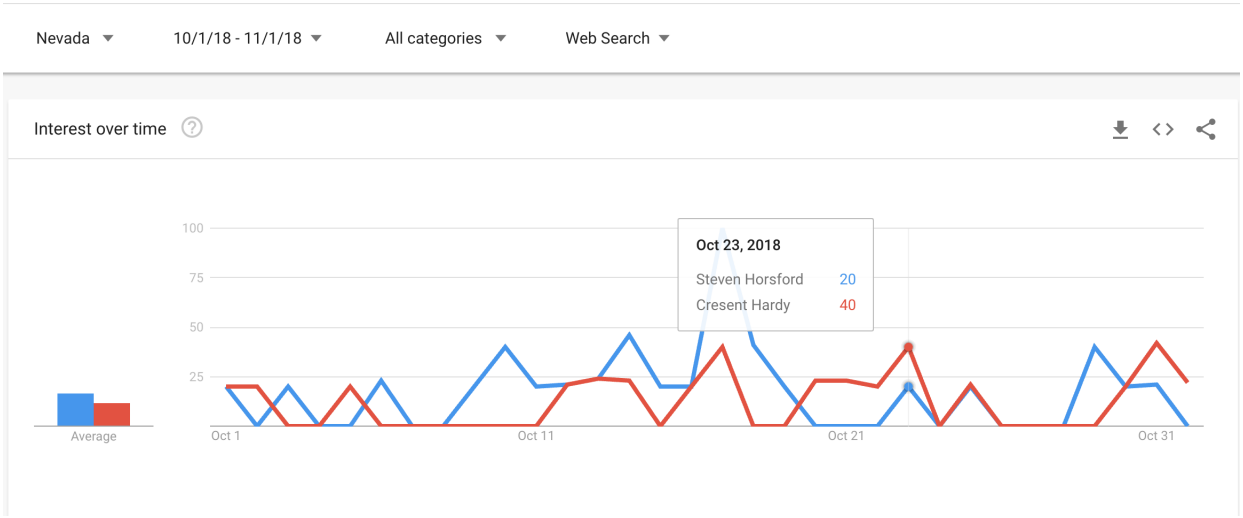

Figure 2: Google Trends Nevada District 4

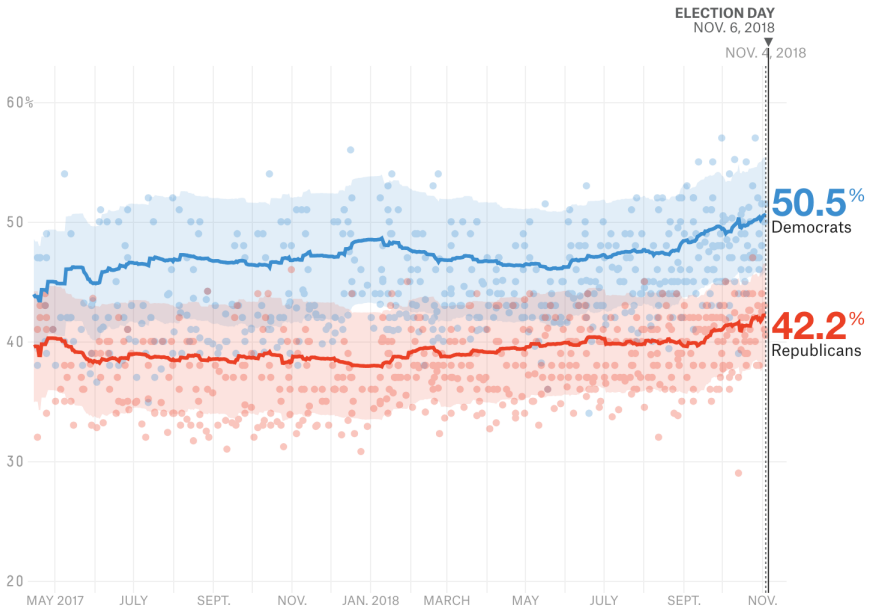

Figure 3: FiveThirtyEight General Polling Data

Today

		DATES	POLLSTER	SAMPLE	RESULT				NET RESULT
U.S. House	WI-6	OCT 29-NOV 3	JMC Analytics / Bold Blue Campaigns	500 LV	Kohl	33%	61%	Grothman	Grothman +28
	OH-7	OCT 31-NOV 1	C- Optimus	800 LV	Harbaugh	36%	55%	Gibbs	Gibbs +18
	AZ-1	OCT 31-NOV 1	C- Optimus	756 LV	O'Halleran	48%	45%	Rogers	O'Halleran +2
	NM-2	OCT 26-NOV 1	A Research & Polling, Inc.	413 LV	Small	45%	46%	Herrell	Herrell +1
	NM-1	OCT 26-NOV 1	A Research & Polling, Inc.	419 LV	Haaland	50%	38%	Arnold-Jones	Haaland +12

KEY A = ADULTS RV = REGISTERED VOTERS V = VOTERS LV = LIKELY VOTERS

Yesterday

U.S. House	KY-6	NOV 1-4	A Siena College/New York Times	438 RV	McGrath	38%	47%	Barr	Barr +9
	KY-6	NOV 1-4	A Siena College/New York Times	438 LV	McGrath	44%	44%	Barr	EVEN
	IA-4	OCT 31-NOV 4	A Siena College/New York Times	423 RV	Scholten	40%	45%	King	King +5
	IA-4	OCT 31-NOV 4	A Siena College/New York Times	423 LV	Scholten	42%	47%	King	King +5
	IL-14	OCT 31-NOV 4	A Siena College/New York Times	428 RV	Underwood	45%	39%	Hultgren	Underwood +6
	IL-14	OCT 31-NOV 4	A Siena College/New York Times	428 LV	Underwood	49%	43%	Hultgren	Underwood +5
	CO-3	OCT 27-NOV 2	JMC Analytics / Bold Blue Campaigns	500 LV	Mitsch Bush	41%	46%	Tipton	Tipton +5

Figure 4: FiveThirtyEight District Polling Data

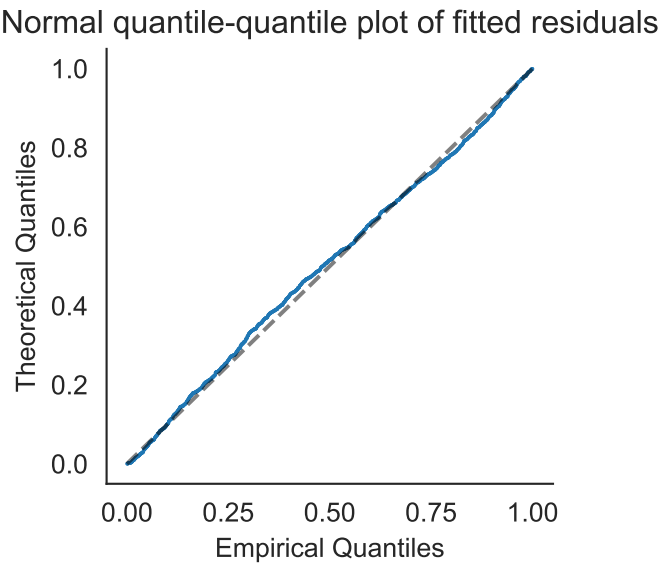

Figure 5: Normal Q-Q plot of residuals in fitting (Estimation for β_0)

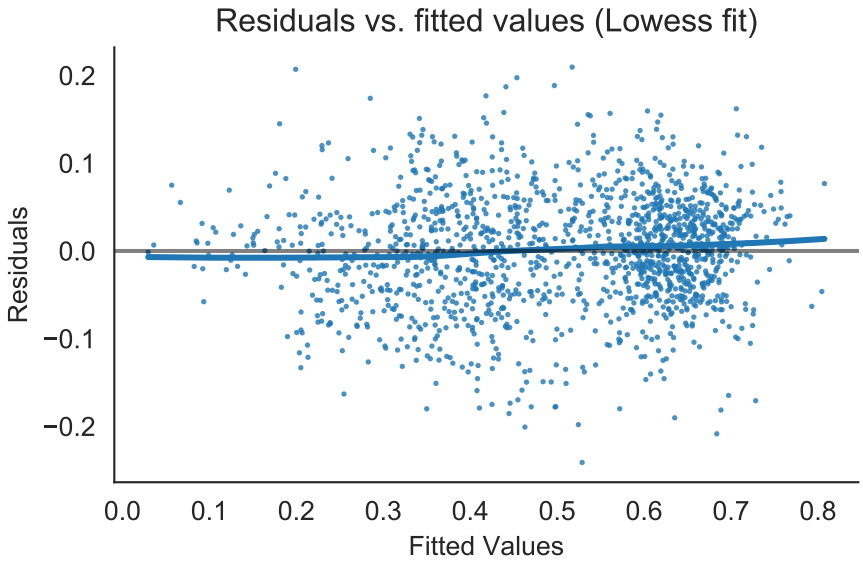

Figure 6: Regression diagnostics in fitting (Estimation for β_0)

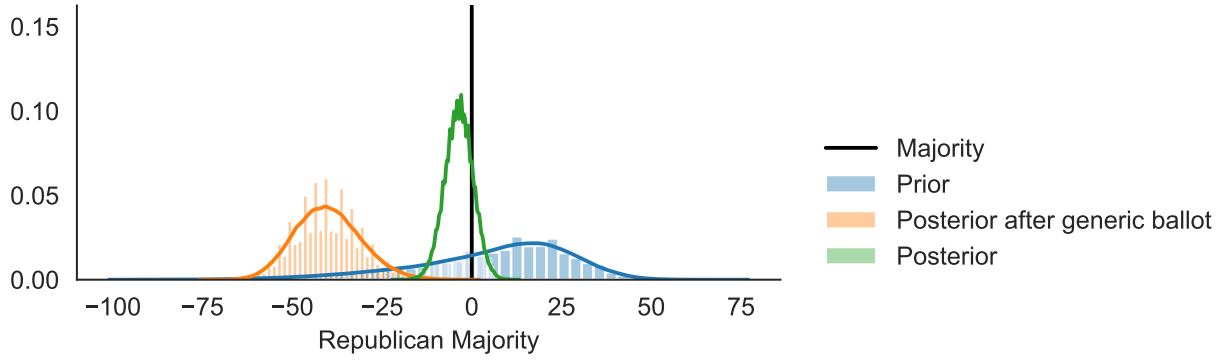

Figure 7: Projected Republican seat distribution from prior, intermediate posterior, and full posterior with fully general Σ_0

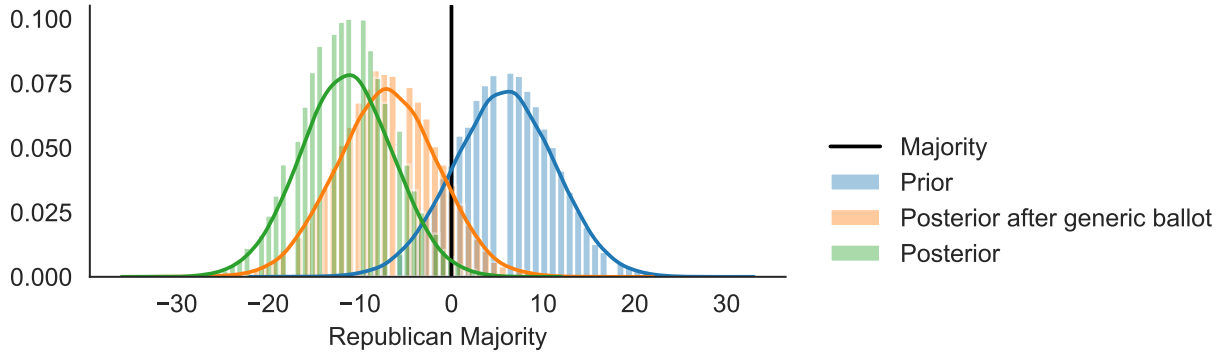

Figure 8: Projected Republican seat distribution from prior, intermediate posterior, and full posterior with diagonal Σ_0

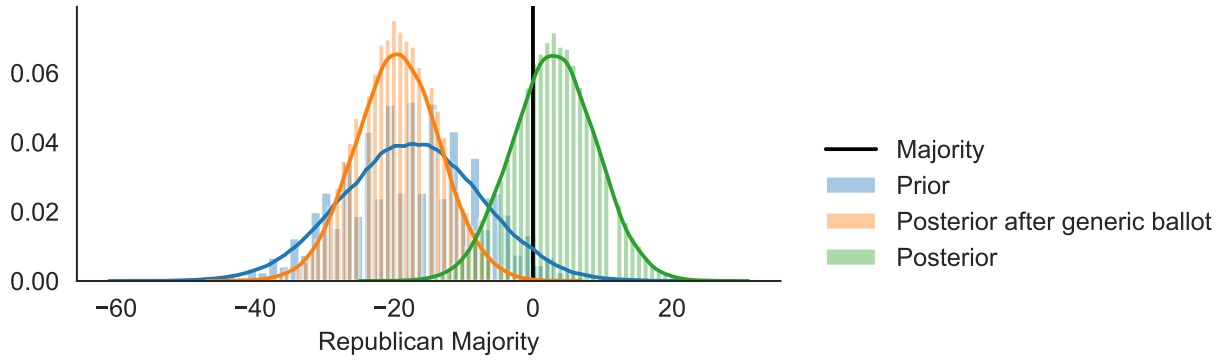

Figure 9: Projected Republican seat distribution from prior, intermediate posterior, and full posterior, if we use a $\mathcal{N}(0.5\mathbf{1}, 10I)$ prior for the contested races. It may be odd that the prior is not centered at zero. The reason is that there are more races in which the Democrat is uncontested than those in which the Republican is. It may also be odd that the full posterior shifts the distribution to the right, unlike in Figure 7. The reason is that the district-wide polls shrink variance dramatically, and if the *polled* districts are disproportionately Republican-leaning, then among the districts we are confident about, more districts are Republican-leaning. This would shift the seat distribution to the right, even though more values of posterior mean are less than 0.5 (Democratic majority), since more Democratic districts have large variance.

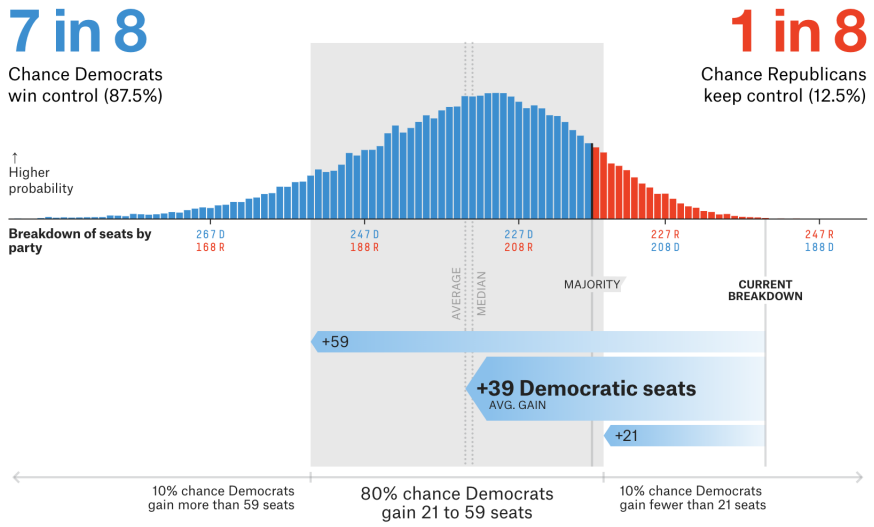

Figure 10: FiveThirtyEight Forecast for House on Nov. 5, 2018

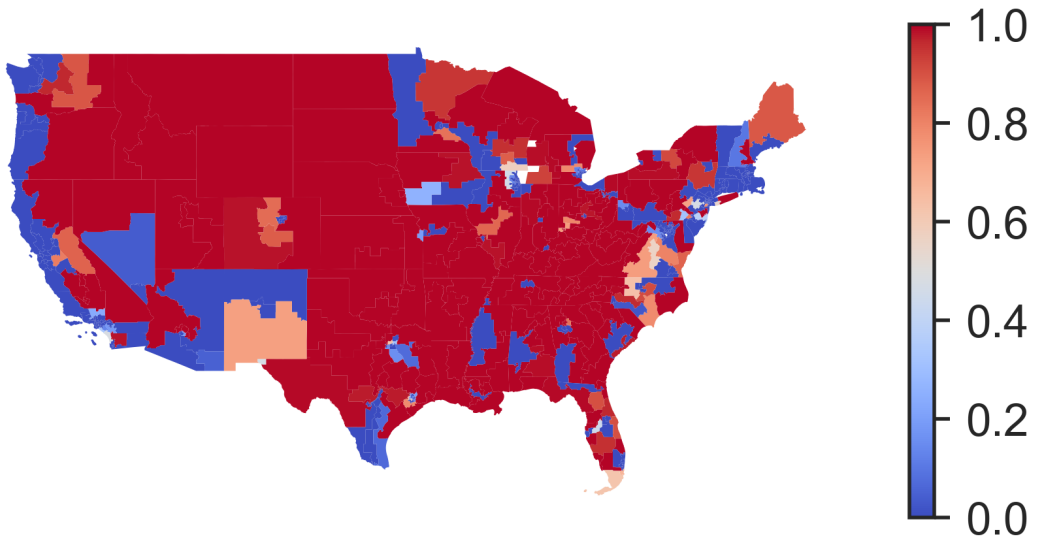

Figure 11: Projected marginal probabilities of Republican victory

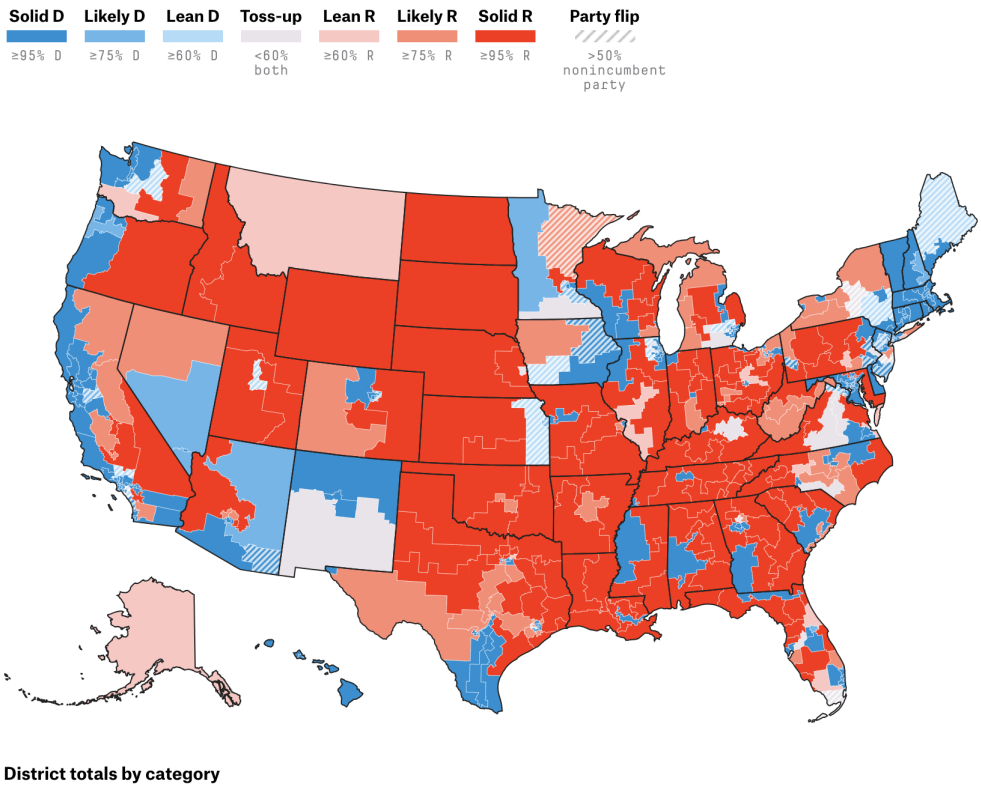

Figure 12: District-level forecast by FiveThirtyEight