Poll-Based Conjugate Prior Models for the Prediction of United States Presidential Elections

Brittany Alexander[1*] and Leif Ellingson[2 †]

Abstract

A previous Bayesian model used to predict the 2008, 2012, and 2016 United States Presidential Elections using only poll data resulted in nearly identical electoral college predictions to FiveThirtyEight, and 95.329% relative accuracy to the FiveThirtyEight Polls Plus model in terms of root mean square error of the predictions of the two major candidates. The previous model used poll data from either another single similar state or national polls to create prior distributions and used the MLE estimators to fit the model. We present new models with minor differences that are used on the same data used in the previous model. The new models now pool the polls together from other states in the regions and uses the pooled estimates as the prior instead of relying on poll data from one state. The new models compare the beta and Gaussian conjugate prior and use three different methods to reassign undecided voters, and either updates iteratively or pools the polls together and performs the calculation once. We also provide a variety of models to serve as comparison for the model's accuracy such as a noninformative model, and polls only model.

Key Words: political polling, conjugate priors, election prediction, American elections

1. Introduction

Every four years the United States holds an election to decide the president. The Electoral College makes the final decision with electors usually bound to support the winner of their state. The popular vote has been successfully modeled for decades using fundamental models based on political and economic data as discussed in Gelman & King (1993). However, the prediction of state-level support in American presidential elections is difficult, as discussed in Hummel & Rothschild (2014), and Gelman & King (1993), and Lock & Gelman (2010). Numerous models attempt to predict Presidential elections on the state level using fundamental models, poll data or both. As a part of a previous project, in Alexander (2019), we created a model that used poll data from other states as the prior to perform a Bayesian analysis of poll data that performed comparably to the FiveThirtyEight Polls Plus model, a popular model used to predict the 2008, 2012, and 2016 elections.

This previous model made a number of simplifications and assumptions that

*research@balexanderstatistics.com
†leif.ellingson@ttu.edu

made the initial analysis more tractable, but also did not reflect the realities of elections as well as it could have. For instance, the model divided the country into five categories (Western blue states, Midwestern red states, Southern red states, Northern blue states, and swing states) and each category had a different prior. It then calculated the maximum likelihood estimators for the variance and the mean of voter support for candidates in each state. The model combined these estimates and the information from Gaussian conjugate priors to predict the support for each candidate in every state. It was primarily designed to prospectively predict the 2016 election and, as such, defined a swing state as "a state were multiple candidates had a reasonable chance of winning the state." This definition did not translate well for retrospective studies on the 2008 and 2012 elections and likely would not be suitable for analyzing and predicting future elections either. The limited number of categories meant that some states were not that similar to the state used to create the prior. The model attempted to predict results for minor candidates in 2016 because of an assumption, that later was determined to be unfounded, that minor candidates would perform significantly better in 2016 than in previous years. The model was incapable of properly capturing minor candidate support since the poll data on minor candidates was more limited than expected. It did not normalize the proportions of voter support to sum to 1 until after results for all candidates were calculated, which had the effect of artificially increasing the variance of the polls compared to what would be obtained if polls were normalized before the analysis.

Despite the above issues, the previous model showed promise. Overall, in terms of root mean square error, it was 95.329% as accurate as the FiveThirtyEight Polls Plus Model at predicting the relative votes of the two major candidates. While it did not perform as well in swing states, where it was just 68.72% as accurate as the FiveThirtyEight Poll Plus model at predicting the relative votes of the two major candidates. Our hope was that if the model was modified so that the required assumptions could more closely match reality, then even merely maintaining the predictive performance of the original model would provide a substantial improvement to the initial results.

As such, in this paper, we present new Bayesian models that we developed to build upon the successes of the earlier model while trying to improve upon its limitations. Our principal goal is therefore to see if these changes have indeed improved the results. To do this, we study the effects that changing the form of the prior distributions, the methods used to handle undecided voters and supporters of minor candidates, and calculation methods have on the viability of using poll data from either other states or national polls in the construction of the priors.

More specifically, this study furthers the research into using poll data to construct prior distributions in a Bayesian analysis to predict the proportion of support for major candidates in an American Presidential election. Despite conjugate priors limiting model flexibility, we continue to use them here due to their computational efficiency, which is helpful in providing near-real-time prospective predictions for an election. The models are built upon a combination of ideas from the previous model and the one discussed in Christensen and Florence (2008). We explore different methods for incorporating poll data into the priors, focusing on using polls from all other states in a state's category rather than just a single representative. Ultimately, we examine twelve similar models that use combinations of two different conjugate priors (Gaussian and Beta), three ways to reassign undecided voters (pro-

portionally, based on past vote, or 50-50), two ways to make the calculation (once or iteratively), and two ways to examine the data (in terms of people or in terms of polls). We also created nine models for evaluating the benefits of the empirical Bayesian framework used in the twelve primary models. In an effort to avoid some of the pitfalls of the previous approach, this study focuses exclusively on predicting support for the two major candidates for the 2008, 2012, and 2016 elections. The model ignores independent and third-party candidates, who typically are a tiny proportion of the overall vote. As such, we estimate the proportion of Democratic support in a particular state in a particular year of the two-party vote and then use this to estimate the proportion of Republican support.

This paper summarizes the results of this study and is organized in the following manner. Section 2 presents the methodology used for all models, Section 3 provides the specific methodology for the 21 models, Section 4 presents the results, and Section 5 is a discussion of the study.

2. General Methodology

2.1 Data Description

This model uses the data from Huffington Post's Pollster previously collected for use in an earlier Bayesian model to predict American presidential elections using poll data. Huffington Post used to provide CSV files of poll data on specific races (i.e., 2016 Presidential General Election in Alabama). They still provide the same data but have now shifted the data format to TSV files.

The following are descriptions of each item in the data.

- RepublicanName: the percent of responses for the Republican candidate in that poll,
- DemocratName: the percent of responses for the Democrat candidate in that poll,
- Undecided: the percentage of undecided voters,
- Other: the percent of responses for other candidates (omitted in select states),
- poll_id: the id number assigned by Pollster to that poll,
- pollster: the name of the polling agency,
- start_date: the start date of the poll,
- end_date: the end date of the poll,
- sample_subpopulation: indicates whether the poll was of registered voters or likely voters,
- sample_size: the sample size of the poll,
- mode: indicates the method that poll was collected (internet, live phone, etc.),
- partisanship: if a partisan group sponsored the poll, and

- partisan_affiliation: indicates the partisan stance of the group that sponsored the poll.

2.2 Data Inclusion Criteria

The data used in the model are from polls conducted between July 1st and the Friday before the election. The reasoning for this is that, after July 1st, the winner of the Democratic and Republican nomination process is clear in these three elections. This effectively excludes polls conducted during the nomination process. Around this point in time, the polls tend to produce data of higher quality since the number of undecided voters decreases and polling become more frequent. The deadline for polls to be conducted by the Friday before the election helps create a group of polls that would most likely have been available for use in making a prospective prediction of the election. The time it takes for a poll to be released varies, but is usually a few business days after the poll ends.

2.3 Prior Specification

The 50 states and Washington DC were put into groups based on the average margins of victory over the past four elections before the election being predicted. We define the margin of victory as the difference between the percent of the vote cast for the Democratic candidate and the Republican candidate. As such, a positive margin indicates that the Democratic candidate won. Then the states are divided into the following groups: Very Strong Democratic States (margin $>$ .2), Strong Democratic States (0.1 $<$ margin $<$ 0.2), Lean Democratic States (0.025 $<$ margin $<$ 0.1), Toss Ups (-0.025 $<$ margin $<$ 0.025), Lean Republican States (-.1 $<$ margin $<$ -0.025), Strong Republican States (-.2 $<$ margin $<$-.1), and Very Strong Republican States (margin $<$ -.2). The model then pools the polls from the other states by taking the average and sample standard deviation of the polls in the subgroup and uses that pooled information to make the priors.

2.4 States without Data

In 2012, Alaska, Delaware, and Wyoming did not have any polls tracked by either Real Clear Politics or Pollster. Since there is no data on these states, the pooled estimators of the poll data from other states in that region are used as the prediction for that state. As state-level polling continues to expand, it might become rarer for states to have no poll data, even if they are extremely partisan or have small populations.

2.5 Estimators

Two different sets of estimators were used to estimate the mean and variance of the distributions. Some models used the normal approximation to the binomial distribution which is defined below:

Normal Approximation to the Binomial Distribution using proportions
Let x_c be the responses for a candidate in a poll, and let N represent the total

number of responses in a poll.

$$\mu = \hat{p} = \frac{x_c}{N}, \qquad \sigma^2 = \sqrt{\frac{(1-\hat{p})\hat{p}}{n}} \tag{1}$$

Other models used the Maximum Likelihood Estimators for the mean and the sample variance are defined below:

Maximum Likelihood Estimators for σ^2 and μ

$$\hat{\mu} = \frac{1}{n}(\sum_{i=1}^{n}(X_i)) \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\sum_{i=1}^{n}(X_i - \bar{X})^2) \tag{3}$$

For the beta model the parameters α, β for the prior distribution were fit using the following formula.

Parameter estimation of the Prior Beta Distribution

Given the estimated mean μ_c and variance $\hat{\sigma_c}$ in equations (2) and (3)

$$\hat{\alpha} = (\frac{1-\mu_c}{\hat{\sigma_c}})\mu_c^2 \tag{4}$$

$$\hat{\beta} = \hat{\alpha}(1/\mu_c - 1) \tag{5}$$

2.6 Minor Candidates

State-level polls often ignore or underestimate minor candidates, which are candidates who receive less than 5% of the popular vote. Minor candidates, thusly by nature, comprise small proportions of the votes. Other than Gary Johnson, who received 3.28% of the popular vote, and Jill Stein, who received 1.07% of the popular vote in 2016, all other minor candidates in 2008, 2012, 2016 received less than 1% of the popular vote. Since the winner of the state decides most of the electors in the electoral college, minor candidates do not have a large role in deciding the winner of the election outside of possibly lowering the number of votes for a major candidate. Additionally, poll data on minor candidates is limited, so it is difficult to build an exclusively poll-based model to predict minor candidates. Furthermore, a model for just the two major candidates allows us to model support for the Democratic candidate using the binomial distribution and then directly use that to obtain results for the Republican candidate. For these reasons, this study focuses on the two major candidates and normalizes the election results proportionally so that the new proportions of Republican and Democratic candidate support sum to 1.

3. Model Specification

3.1 Model List

Twenty-one models were used in the experiment, representing seven calculation methods repeated three times using the three different normalization methods. The seven model types will be referred to as follows. The Bayesian models all use poll

data from other states in the region in prior construction. The tested Bayesian models are: **the Beta model**, which uses the conjugate prior for binomial data, **the Gaussian Iterative model**, which is an iterative Gaussian model, **the Gaussian People model**, which is a Gaussian model that pools the responses into a giant poll, and **the Gaussian Polls model**, which is a Gaussian model that averages the polls. The accuracy comparison models, which are used only to compare the other models to, are: **Polls Only model**, which is the model that simply takes the average of the polls, **Prior Polls model**, which simply finds the average of the polls in the prior region, and **the NI (noninformative) model**, which uses a noninformative Beta(1,1) prior.

3.2 Conjugate Prior

For the tested Bayesian models, either the beta conjugate prior or the Gaussian conjugate prior was used in the analysis, depending on the form of the data used to construct the likelihood function. We assume that the data is identically and independently distributed to simplify the calculation. The data is naturally binomial, which gives rise to the beta conjugate prior. However, the data was treated as Gaussian in some of the models for the purpose of comparing to the previous model and to consider the normal approximation of the binomial distribution, thus necessitating the use of the Gaussian conjugate prior in such scenarios. In most cases, the Gaussian-based models used the normal approximation to the binomial distribution, except pooled Gaussian models using the number of polls to determine the number of observations. The sample sizes of these polls are relatively large and are usually comprised of at least a few hundred people, with some polls having more than a thousand respondents. Additionally, in most cases, the proportions are in the range of .4 and .6. As such, although no transformation of the data was made, the area of the normal curve smaller than 0 or larger than 1 has an infinitesimal probability due to the fact that the standard deviation was usually less than 0.05. The poll data may not be perfectly normal, but we assume it is normal enough for the analysis to work. Below, are the conjugate priors used in the models.

Gaussian Conjugate Prior

In the normal models, given the sample mean $\bar{p}_{c_0}$ and variance $\sigma^2_{c_0}$ of the polls of the state used as the prior for a candidate c, and the sample mean $\bar{p}_c$, number of polls n, and variance σ^2_c of the polls from the state being analyzed for a candidate c, the posterior distribution of support for a candidate c is a Normal distribution with updated mean and variance parameters can be calculated as follows:

$$\mu_c = \frac{\hat{\sigma}^2}{n\hat{\sigma}^2_{c_0} + \hat{\sigma}^2_c}\bar{p}_{c_0} + \frac{n\hat{\sigma}^2_{c_0}}{n\hat{\sigma}^2_{c_0} + \hat{\sigma}^2_c\bar{p}}, \qquad \sigma^2_c = (\frac{1}{\hat{\sigma}^2_{c_0}} + \frac{n}{\hat{\sigma}^2_c})^{-1} \tag{6}$$

Beta-Binomial Conjugate Prior

For the binomial models we assume a binomial likelihood, the posterior distribution of support given x_c responses for the candidate being predicted and N responses total is

$$beta(\alpha_{post} = \alpha + x_c,\ \beta_{post} = \beta + N - x_c). \tag{7}$$

Recall that the mean and variance of a beta random variable are:

$$\mu = \frac{\alpha}{\alpha + \beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2} \tag{8}$$

3.3 Normalization

Three different methods were used to reassign undecided and minor candidate voters: proportionally (Prop) to the poll data, proportionally based on prior election (Past Vote or PV) results, or splitting the undecided voters evenly between the candidates. Proportional reassignment took the support of minor candidates or undecided voters and divided that support proportionally based on the poll data. Past vote reassignment took the support of minor candidates or undecided voters and divided that support based on the past vote. Splitting the undecided voters between the two candidates is done similarly to the past vote method with r_{old} and d_{old} equal to 0.5. Below are the formulas used to normalize the data.

Proportional Normalization Calculation

$$r_{new} = \frac{r_{old}}{r_{old} + d_{old}} \tag{9}$$

$$d_{new} = \frac{d_{old}}{r_{old} + d_{old}}, \tag{10}$$

where r_{new} and d_{new} are the normalized values for the republican, and democratic candidates, respectively, and r_{old}, and d_{old} are the original predictions for the republican and democratic candidates, respectively.

Past Vote Normalization Calculation

$$norm = r_{old} + d_{old} \tag{11}$$

$$r_{new} = r_{old} + (1 - norm) * r_{vote} \tag{12}$$

$$d_{new} = d_{old} + (1 - norm) * d_{vote}, \tag{13}$$

where r_{vote}, and d_{vote}, are the past two-party election results from the state, and r_{new}, and d_{new} are the normalized values for the republican, and democratic candidates, respectively, and r_{old}, and d_{old} are the original predictions for the republican and democratic candidates, respectively.

3.4 Pooling vs. Iterative

Two different ways of evaluating all the poll data were used. The poll data was either pooled together and the conjugate prior calculation was performed once, or the conjugate prior calculation was performed iteratively. Due to the nature of the beta conjugate prior, the pooled version and iterative version are equivalent, and the pooled version was used in the calculations to cut down on calculation time. For the Gaussian-based models, there were two distinct groups of models: people-based or poll-based. The Gaussian people-based models used the binomial approximation to the normal distribution and had both pooled (Gaussian People) and iterative (Gaussian Iterative) versions. The Gaussian poll-based models were strictly pooled models and used the mean and variance of the observed polls inside the conjugate prior calculation.

3.5 Pooled Calculation

Based on the percentage support for the Democratic candidate after the reassignment of undecided voters and the sample size, the number of people who supported the democratic candidate was synthesized by multiplying the Democratic candidate's support and the sample size and rounding that number to the nearest integer. In the case of a number where the decimal portion was exactly 0.5 the program rounded to the nearest even integer. The total number of people supporting the democratic candidate and the overall sample size was calculated and used to estimate the proportion of people supporting the democratic candidate for pooled models. The overall counts and estimated proportions were then used in the respective conjugate priors.

4. Results

There are a few main methods to evaluate Presidential Election forecasts. One method calculates the percentage of races called correctly, but this method is not helpful at comparing various forecasts since virtually all models predict approximately the same winners. Included below in Table 1, is the number of correctly called states in the 2008, 2012, and 2016 elections, along with the overall percentage accuracy of all the models and selected external models is listed.

As shown below, all models perform similarly at predicting the winners of the states. The models all missed Ohio in 2008, and Florida, Michigan, North Carolina, Pennsylvania, and Wisconsin in 2016. In 2012, the Iterative models missed Florida by less than a tenth of a percentage point. The new models, with the exception of the Prior Only model, perform about as well as the previous model and the FiveThirtyEight model at predicting the winner in states. These results are promising because the new models are incredibly quick computations.

Table 1: Percentage of Election Winners Called Correctly

Model	2008	2012	2016	% Accuracy
Prior Proportional	46	47	44	90.196
Prior Past Vote	45	48	44	88.235
Prior 50-50	46	47	44	90.196
Polls Proportional	50	51	46	96.078
Polls Past Vote	51	51	46	95.425
Polls 50-50	50	51	46	96.078
Beta Proportional	50	51	46	96.078
Beta Past Vote	51	51	46	96.078
Beta 50-50	50	51	46	96.078
Iterative Proportional	48	51	46	95.425
Iterative Past Vote	49	51	46	95.425
Iterative 50-50	49	51	46	95.425
Gaussian People Proportional	50	51	46	96.078
Gaussian People Past Vote	51	51	46	96.078
Gaussian People 50-50	50	51	46	96.078
Gaussian Polls Proportional	50	51	46	96.078
Gaussian Polls Past Vote	51	51	46	95.425
Gaussian Polls 50-50	50	51	46	96.078
Noninformative Proportional	50	51	46	95.425
Noninformative Past Vote	51	51	45	95.425
Noninformative 50-50	50	50	46	95.425
Real Clear Politics[1]	49	50	47	95.425
Princeton Election Consortium[2]	50	50	46	95.425
Five Thirty Eight (Polls Plus)[3]	50	51	46	96.078
PredictWise (Fundamental)[4]	N/A	50	46	94.118
Sabato's Crystal Ball [5]	51	49	46	95.425
Previous Model (Alexander 2019)	50	51	45	95.425

[1] RealClearPolitics.com

[2] Princeton Election Consortium: election.princeton.edu/

[3] FiveThirtyEight Polls Plus Model: fivethirtyeight.com

[4] Predict Wise: predictwise.com/

[5] http://crystalball.centerforpolitics.org/crystalball/2020-president/

A second measure of accuracy is to calculate the average error and root mean square error of the predicted elections results compared to the results of the elections. Other measures for evaluating probabilistic predictions, such as Brier scores, are not relevant in this case since the models were not designed to make probabilistic predictions. Table 2 provides the Average Error (AE) and Root Mean Square Error (RMSE) of the Proportional Models. Table 2 includes the AE and RMSE for each election for the tested models, the previous model, and the FiveThirtyEight model (538), as well as the average AE and RMSE across all elections, and the relative accuracy for each of the models to the FiveThirtyEight (538), Polls Only, and Previous Model. Table 3 provides the Average Error and Root Mean Square Error of all the different normalization methods for the Iterative and Polls Only Model.

Additional maps and tables can be found in the Appendix.

Table 2: Average Error and Root Mean Square Error of Proportional Models

	Prior	Polls	Beta	GI[1]	GPe[2]	GPo[3]	NI[4]	PM[5]	538[6]
2008 AE	5.021	2.571	2.698	2.318	2.522	2.663	2.525	2.496	2.045
2008 RMSE	7.324	3.151	3.528	3.382	3.099	3.338	3.106	2.985	3.196
2012 AE	4.563	1.939	2.038	1.946	1.894	1.995	1.894	1.831	1.602
2012 RMSE	6.839	2.571	2.781	2.670	2.562	2.64	2.563	2.367	1.977
2016 AE	5.077	2.952	3.073	3.004	3.047	3.009	3.047	3.228	3.049
2016 RMSE	7.364	3.539	3.654	3.568	3.614	3.634	3.614	3.960	3.813
Average AE	4.887	2.487	2.603	2.423	2.487	2.556	2.489	2.518	2.232
Average RMSE	7.175	3.087	3.321	3.207	3.092	3.205	3.094	3.104	2.995
538 AE Comparison	0.457	0.897	0.858	0.921	0.898	0.873	0.897	0.886	1
538 RMSE Comparison	0.417	0.970	0.902	0.934	0.969	0.934	0.968	0.965	1
Polls AE Comparison	0.509	1	0.956	1.027	1	0.973	1	0.988	1.114
Polls RMSE Comparison	0.430	1	0.930	0.963	0.999	0.963	0.998	0.995	1.031
PM AE Comparison	0.515	1.012	0.968	1.039	1.013	0.985	1.011	1	1.128
PM AE Comparison	0.433	1.005	0.935	0.968	1.004	0.968	1.003	1	1.036

[1] Gaussian Iterative
[2] Gaussian People
[3] Gaussian Polls
[4] Noninformative
[5] Previous Model
[6] FiveThirtyEight Polls Plus Model: fivethirtyeight.com

Table 3: Comparison of Normalization Methods

	Polls Prop	Polls PV[1]	Polls 50[2]	GI[3] Prop	GI PV	GI 50
2008 AE	2.57125	2.69039	2.73438	2.31835	2.1815	2.19475
2008 RMSE	3.15145	3.27921	3.45886	3.38187	3.02187	3.15738
2012 AE	1.93866	1.92149	2.19725	1.94647	1.99711	1.99794
2012 RMSE	2.57091	2.55879	2.84355	2.67003	2.75636	2.72241
2016 AE	2.95246	2.80921	3.48908	3.00391	2.83348	3.13048
2016 RMSE	3.53912	3.43298	4.30123	3.56837	3.43685	3.66724
Average AE	2.48746	2.4737	2.8069	2.42291	2.33736	2.44106
Average RMSE	3.08716	3.09033	3.53455	3.20676	3.07169	3.18234

[1] Past Vote
[2] 50-50
[3] Gaussian Iterative

Below, in Figure 1, a choropleth map of the average error of the Gaussian Iterative model is displayed, in Figure 2, a choropleth of the average error of the Beta model is displayed, and in Figure 3, a choropleth map of the average error of the Polls only model is displayed. These plots show on a state-level the average error of the three elections predicted. Additional tables and maps can be found in the Appendix.

Figure 1:

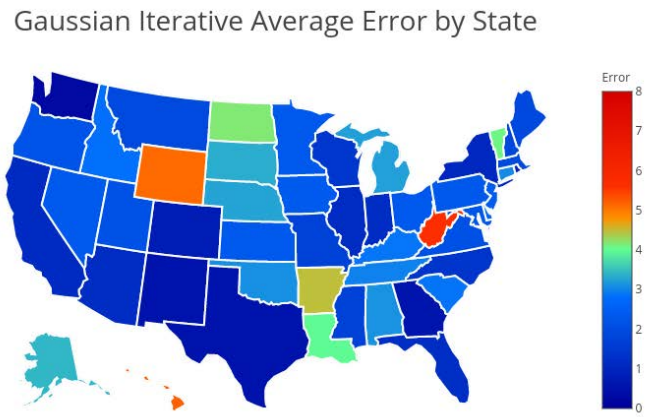


Figure 2:
Beta Average Error by State

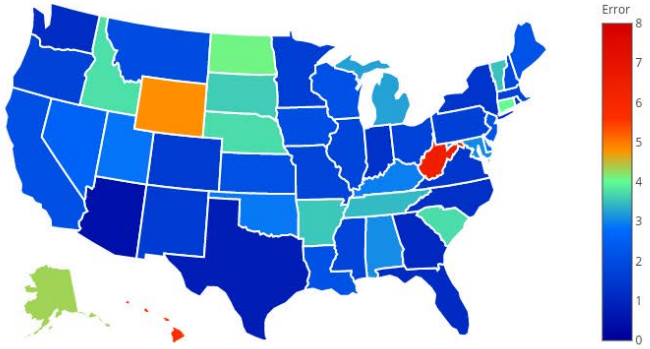
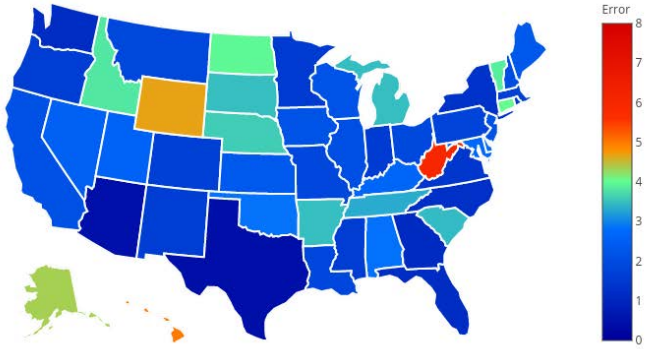


Figure 3:
Polls Only Average Error by State



It is important to acknowledge that while this model consisted of 3,000 individual predictions over 51 areas and 3 elections, this study doesn't have the power to definitely distinguish which model is the best. Many of the tested models were not meaningfully different in their predictions or their performance. The assumptions and the ease of use of a model also have to be considered.

4.1 Comparison of Bayesian vs. Non-Bayesian Methods

The Iterative Gaussian Past Vote model is the best performing model, followed by the Iterative Gaussian Proportional model. Besides the Iterative Gaussian model and the Gaussian People model, the other models do not outperform the Polls Only model, but they are close. The Bayesian methods typically work well except for certain states where the model underpredicts the lead of the winning candidate. In this context, it is probably better to predict a race as more competitive than less competitive due to the known underperformance of polls from their margin of error. In the worst performing states, the errors are generally at least somewhat consistent and can likely be adjusted for in future research to reduce the error even more.

4.2 Comparison of Gaussian vs. Beta

The Beta model did not perform well, likely due to large initial values of a and b derived from the prior. Since the prior does not do well in about 15 states, the higher weight placed on prior information hurts the prediction. The fact that the Beta model performs worse than the noninformative prior also shows issues with the Beta model. However, other model specifications using beta and binomial distributions might provide a better balance between the weight of the prior in the model. The Gaussian model performs well, but it is clear that in the context of this problem, the data might not be adequately approximated by a Gaussian distribution due to the heavy tails of poll data that are observed. Since the model is not concerned with probabilistic predictions, the treatment of the Gaussian models of the variance as a fixed quantity instead of placing a gamma prior on the variance has no effect on the predicted mean.

4.3 Comparison of Pooling vs. Iterative

The iterative analysis best handles a variety of situations. It convergences quickly to the distribution of the data since the weight of the prior is essentially a power series while the non-iterative methods converge less quickly. This quick convergence is ideal for the high polling volume states that are different from the rest of the region like Arizona, Texas, and New York. The iterative analysis also has the advantage of beating the Polls Only model in most cases. The iterative analysis also has the benefit of more closely mirroring the polls than the prior which is helpful in most cases because the prior does not always perform well. The assumptions of the iterative analysis also make more sense because, within a poll, people should be relatively independent of each other and polls should be relatively independent of each other. However, the Beta and Gaussian People models methods assume that all people surveyed are independent, which is untrue because people often participate in more than one poll.

4.4 Comparison of Normalization Methods

The different normalization methods each varied in their performance. The proportional and past vote normalization were, on average, within a few hundredths of a point. Sometimes proportional normalization was better and other times past vote normalization was better. This suggests that we cannot distinguish between the two methods. From a practical perspective, though, the implementation of a past vote method requires extra effort. Proportional normalization also appears to have theoretical advantages. Given the fluctuations in vote share over time, the past vote method uses older data and is also more complicated since it requires compiling information about the past vote, while the proportional and 50-50 methods do not require that information. The proportional method is also more intuitive and makes fewer assumptions since it is essentially equivalent to removing the undecided voters. Recent work in Bon, Ballard, & Baffour (2019) proposes a new method of handling undecided voters that is more accurate than the other normalizations methods discussed in this article.

4.5 Comparison of the Previous Model and The New Models

The new models in this article attempt to address some of the flaws in the original model. The original prior, which was based on regional categories and used data from one state, was not always accurate. Initially, the second attempt of defining a prior refined the regional categories into smaller groups and used the polls from all the states in that region.

Surprisingly, this prior choice applied to the models in this article created worse results than the original model. Because, the second attempt did not work well, the prior in this article was developed and it resulted in an improvement. The main benefit of the new prior elicitation is that it removes almost all subjectivity with the exception of the choice of cutoff points, which is partially arbitrary. These new models also normalized the polls to sum to one before the analysis instead of after.

The changes in the model proved to be beneficial and resulted in all of the Gaussian-based models with proportional or past vote normalization to have a slightly lower root mean square error than the original model. If we ignore a few states in which the Gaussian Iterative model performed worse in part because of the greater weight placed on polls that happened to be inaccurate, the error reduction was relatively uniform, regardless of the quantity or quality of the polls.

5. Discussion

5.1 Implications of this Study

There is still a need for examining assumptions and building better models after the failure of models to predict the 2016 election. This project represents one more step forward in the long journey to create a model for American elections that is accurate in determining both the winners of elections and the exact results and that is also fast enough to be used to prospectively predict elections. A better estimation of the variance of the model is also needed so that probabilistic representations of the outcomes can be generated.

5.2 Potential Further Research

The prior polls model helps to show the limitations of the exchangeability between states within a region. Demographic differences likely explain part of this variation. However, these results combined with the state effect analysis done in Gelman (2009) show that voting cannot be entirely explained by demographics and that similar voters in different states respond differently. In other words, similar voters in different states vote slightly differently. If a state level effect could be estimated for a certain state, perhaps poll data from other states could be post-stratified not only for demographic characteristics but also for the state effect. Data from the same election cycle has two distinct advantages over older data because it uses the same candidates and the same political conditions. If the exchangeability problem could be overcome, it could make models more accurate at predicting the exact result and determining the closeness of a race.

An interesting question arising from this study and the previous study is the role that regional effects play in the differences across states. In Gelman (2009, pg. 44), a graph illustrates the difference in state-level and national votes over time by

regions. This graph shows that, while the results varied between states, within a region the changes were relatively similar over time. We hypothesize that regional effects are responsible for some of the variation between states across the country, thus partially justifying the assumption of exchangeability between states. If regional effects could be shown to be a factor in vote determination, it could help to justify the use of regional poll data. Post-stratification of national polls to state polls has been done before in Park, Gelman & Bafumi (2004), but we were unable to find a post-stratification of state polls to use in other states. Future research needs to help examine state and regional effects to better model the cultural and economic conditions that are difficult to incorporate into a model.

The use of fundamental modeling, which uses political and economic data like the previous vote share for a party, change in income, and other data points, has been used before to predict American elections. Two successful models include a state-level fundamental model discussed in Hummel and Rothschild (2014), and a Bayesian model, which uses post-stratification national polls and a fundamental model as the prior. A next step would be to build a hierarchical model that would incorporate both a fundamental model and post-stratified poll results from the states and then use that model to predict the 2020 Presidential Election. However, some data may not be possible to post-stratify because the information on demographic variables may not be available.

Another interesting future application of this model would be the American Senate and House elections. A version of this set of models was used to predict the 2018 Senate elections. States were categorized based on the polling data but there were not formal cutoffs. The Iterative models had an average error of 3.624 and called 34 out of 35 races correctly. The other models performed similarly, and the extension of these models to more Senate and House data would make an interesting project.

5.3 Political Bias

It is important to mention that the primary author's political biases could have subtly affected, while unintentionally, the formation of the methodology and interpretation of the results. The primary author is a conservative independent that voted for an unregistered write-in candidate in the 2016 general election and for Marco Rubio in the 2016 Republican primary. The goal was to remain as objective as possible. It is important to note that nearly all models could be subject to influence by personal biases of the designer(s) of the model.

5.4 Conclusion

This study showed the value of using poll data from other states to form a prior distribution in a Bayesian analysis. While most of new models did not beat the previous model in average error or root mean square error, the results indicated the possible promise of a Gaussian iterative-based approach and proportional normalization. However, the slightly lower accuracy of the new models relative to the previous model may be practically and statistically insignificant. The issues in exchangeability of poll data from other states were identified based on the Prior Polls method. These models also provide a quick and accurate way of determining the winning Presidential candidate on the state and national level. These models also

provide some estimate of how close the race is by approximating the margin between two candidates. In its present form, the iterative Gaussian model can be used to predict elections well. Future research will attempt to incorporate the model with more advanced and established methods of interpreting poll data such as hierarchical modeling, fundamental modeling, and poststratification in hopes to reduce the error even more so that the competitiveness of a race can be better classified.

REFERENCES

Alexander, B. (2019), "A Bayesian Model for the Prediction of United States Presidential Elections, SIAM Undergraduate Research Online, 12.

Bon, J. J., Ballard, T. and Baffour, B. (2019), "Polling bias and undecided voter allocations: US presidential elections," 2004–2016. Journal of the Royal Statistical Society, Series A (Statistics in Society), 182(2): 467–493. http://dx.doi.org/10.1111/rssa.12414

Christensen, W.F., & Florence, L.W.(2008), "Predicting Presidential and Other Multistage Election Outcomes Using State-Level Pre-Election Polls," *American Statistician*, 62:1, 1–10

Gelman, A.(2009). *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*," Princeton, NJ: Princeton University Press.

Gelman, A.,& King, G. (1993), "Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?," *British Journal of Political Science*, 23:04 409–451

Hummel, P., & Rothschild, D.(2014), "Fundamental Models for Forecasting Elections at the State level, *Electoral Studies*," 35, 123–139

Lock, K., & Gelman, A. (2010), "Bayesian Combination of State Polls and Election Forecasts. *Political Analysis*," 18(3), 337–348.

Murphy, K. (2007, October 03), "Conjugate Bayesian analysis of the Gaussian distribution" https://www.cs.ubc.ca/ murphyk/Papers/bayesGauss.pdf

Park, D., Gelman, A., & Bafumi, J.(2004), "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls," *Political Analysis*, 12, 375–385

6. Appendix

6.1 Additional Choropleths

Figure 4: Choropleth of the Average Error of the Gaussian People Model

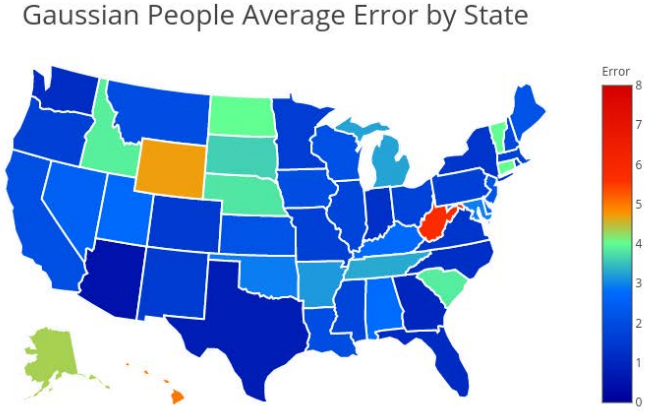


Figure 5: Choropleth of the Average Error of the Gaussian Polls Model

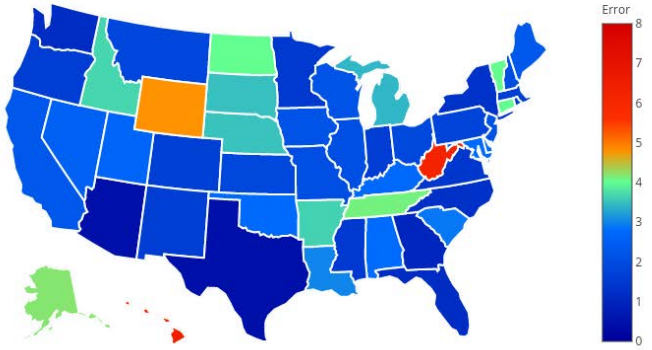


Figure 6:
Noninformative Average Error by State

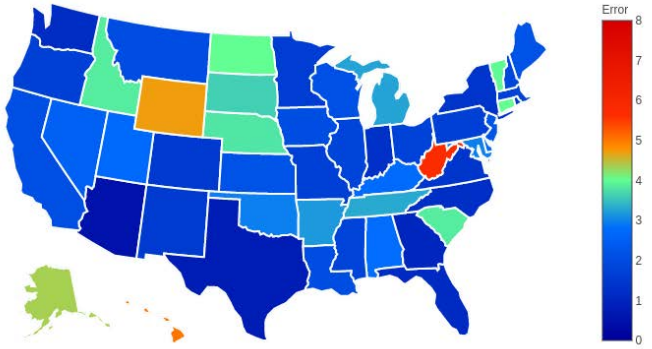
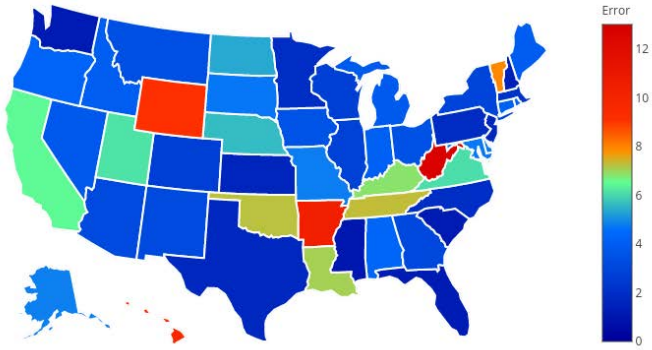


Figure 7:
Prior Only Average Error by State



6.2 Additional Tables

Table 4: Past Vote Average Error (AE) and Root Mean Square Error (RMSE) for Selected Models

	Prior	Polls	Beta	GI[1]	GPe[2]	GPo[3]	NI[4]	PM[5]	538[6]
2008 AE	4.934	2.690	2.812	2.182	2.651	2.781	2.654	2.496	2.045
2008 RMSE	7.015	3.279	3.595	3.022	3.231	3.472	3.237	2.985	3.196
2012 AE	4.425	1.921	2.075	1.997	1.902	1.987	1.905	1.831	1.602
2012 RMSE	6.980	2.559	2.923	2.756	2.561	2.629	2.565	2.367	1.977
2016 AE	4.745	2.809	2.886	2.833	2.861	2.868	2.863	3.227	3.049
2016 RMSE	6.742	3.433	3.503	3.437	3.467	3.510	3.468	3.960	3.813
Average AE	4.701	2.473	2.591	2.337	2.471	2.545	2.474	2.518	2.232
Average RMSE	6.912	3.090	3.340	3.072	3.086	3.204	3.090	3.104	2.995
538 AE Comparison	0.475	0.902	0.861	0.955	0.903	0.877	0.902	0.886	1
Polls AE Comparison	0.433	0.969	0.897	0.975	0.970	0.935	0.969	0.965	1
Polls RMSE Comparison	0.536	1.018	0.972	1.077	1.019	0.989	1.018	1	1.128
Polls AE Comparison	0.449	1.004	0.929	1.010	1.006	0.968	1.004	1	1.036
PM[5] AE Comparison	0.529	1	0.955	1.058	1.001	0.972	1	0.9878	1.114
PM[5] RMSE Comparison	0.4467	1	0.925	1.006	1.001	0.965	1	0.995	1.031

[1] Gaussian Iterative
[2] Gaussian People
[3] Gaussian Polls
[4] Noninformative
[5] Previous Model
[6] FiveThirtyEight Polls Plus Model: fivethirtyeight.com

Table 5: 50-50 Average Error (AE) and Root Mean Square Error (RMSE) for Selected Models

	Prior	Polls	Beta	GI[1]	GPe[2]	GPo[3]	NI[4]	PM[5]	538[6]
2008 AE	5.199	2.734	2.882	2.195	2.693	2.830	2.696	2.496	2.045
2008 RMSE	7.503	3.459	3.917	3.157	3.414	3.674	3.422	2.985	3.196
2012 AE	4.713	2.197	2.445	1.998	2.174	2.291	2.177	1.831	1.602
2012 RMSE	7.001	2.844	3.218	2.722	2.833	2.935	2.839	2.367	1.977
2016 AE	5.258	3.489	3.614	3.130	3.593	3.542	3.595	3.227	3.049
2016 RMSE	7.707	4.301	4.453	3.667	4.406	4.402	4.409	3.960	3.813
Average AE	5.057	2.807	2.980	2.441	2.820	2.888	2.822	2.518	2.232
Average RMSE	7.403	3.535	3.863	3.182	3.551	3.670	3.557	3.104	2.995
538 AE Comparison	0.441	0.795	0.749	0.914	0.792	0.773	0.790	0.886	1
538 RMSE Comparison	0.405	0.847	0.775	0.941	0.843	0.816	0.842	0.965	1
Polls AE Comparison	0.492	1	0.942	1.150	0.995	0.972	0.994	0.988	1.114
Polls RMSE Comparison	0.417	1	0.915	1.1107	0.995	0.963	0.994	0.995	1.031
PM AE Comparison	0.498	0.897	0.845	1.032	0.893	0.872	0.892	1	1.128
PM RMSE Comparison	0.419	0.878	0.804	0.975	0.874	0.846	0.873	1	1.036

[1] Gaussian Iterative
[2] Gaussian People
[3] Gaussian Polls
[4] Noninformative
[5] Previous Model
[6] FiveThirtyEight Polls Plus Model: fivethirtyeight.com