

A Bayesian Model for the Prediction of United States Presidential Elections*

Brittany Alexander[†]

Advisor: Leif Ellingson[‡]

Department of Mathematics and Statistics, Texas Tech University

Abstract

Using a combination of polling data and previous election results, FiveThirtyEight successfully predicted the Electoral College distribution in the presidential election in 2008 with 98% accuracy and in 2012 with 100% accuracy. This study applies a Bayesian analysis of polls, assuming a normal distribution of poll results using a normal conjugate prior. The data were taken from the Huffington Post’s Pollster. States were divided into categories based on past results and current demographics. Each category used a different poll source for the prior. This model was originally used to predict the 2016 election, but later it was applied to the poll data for 2008 and 2012. For 2016, the model had 88% accuracy for the 50 states. For 2008 and 2012, the model had the same Electoral College Prediction as FiveThirtyEight. The method of using state and national polls as a prior in election prediction seems promising and further study is needed.

1 Introduction

A presidential election happens in the United States every four years. The winner of the election is decided by the 538 electors in the Electoral College. Under the current system, most of the states choose their electors based on the winning candidates of the state. Most of the states (like Delaware and West Virginia) have voted for the same party in most recent elections. These states are called “red” if they consistently vote for the Republican and “blue” if they consistently vote for the Democrat. But in some states, like Florida, the winner is not consistent, and the state occasionally “swings” from one party to another. The key to winning a presidential election is winning as many of these “swing states” as possible.

The outcomes of presidential elections have been predicted by political scientists and statisticians for decades. Some of the first models used economic and political data to predict the winner of the popular vote. The challenge is to predict the outcome in swing states, which ultimately win a candidate the election. The quality and quantity of state level polling data has increased over time. All states in the 2016 election had at least 2 polls, while in the 2008 and 2012 election, some

*A earlier version of paper was an Honorable Mention in the Spring 2017 Undergraduate Research Project Competition.

[†]research@balexanderstatistics.com

[‡]leif.ellingson@ttu.edu

states had no scientific polls. Poll data are easily accessible through the various websites. This increase in data has allowed for more accurate predictions of the voting results. Indeed, using a combination of polling data and previous election results, Nate Silver successfully predicted the Electoral College distribution in the presidential election in 2008 with 98% accuracy and in 2012 with 100% accuracy.

Another problem is the limited information on state level support. As discussed in [5] and [7], estimations of state level support between candidates are difficult due to limited polling information on a state by state level. While the number of polls in individual states has increased over time, there is still a lack of data in smaller states. States like West Virginia may only have one or two polls. To help address this problem, states with similar demographics were used as priors for this model.

For purposes of simplicity, the data were assumed to be normally distributed. The normal conjugate prior was used for the means, resulting in a normal posterior. This is obviously not ideal considering the nature of polls, but this method greatly simplified the process. The goal of this study was to test if the inclusion of current poll data from other areas with a Bayesian analysis would help adjust for limited information in states with little poll data.

This study, which was based on methods discussed by [1], applies a Bayesian analysis of polls, assuming a normal posterior using a conjugate prior. While numerous models have been studied as methods to predict presidential elections, the idea of a Bayesian approach using the poll data of other similar states as priors has not been pursued as rigorously. The approach proposed by [1] used previous elections as a prior, but the differences in candidates from different election cycles means that some voters may not vote for the same party as they did in previous elections.

The goal of this study was to test if the inclusion of current poll data from other areas with a Bayesian analysis would help adjust for limited information in states with little poll data. This model was originally used to predict the 2016 election, but later it was applied to the poll data for 2008 and 2012.

The model worked well at predicting the winners of the states and therefore predicting the electoral college results. The model provided close estimates of the final results but there is improvement in creating more precise estimates and improving the predictions of minor candidates. Further research should remove some of the assumptions and develop a more accurate prior distribution by including more polls and adjusting the polls for state-level affects.

This paper summarizes the results of this study and is organized in the following manner. Section 2 presents the methodology used, Section 3 presents the results, and Section 4 is a discussion of the study.

2 Methodology

The guidelines and methodology of the model were set in September 2016 to allow adequate time to prepare the computer programs necessary to run the analysis before the election. The main focus was on the 2016 election, but a later analysis was performed on the 2008 and 2012 elections to get more results on the accuracy of the method.

2.1 Data

During the prediction process, Pollster ([3]), run by the Huffington Post, had poll results available in csv format for 2012 and 2016 in November of 2016. Since Pollster provided the results in an easy to analyze format, they were chosen as the data source for the analysis. Pollster no longer has data in this format but they now instead provide tsv files that can be analyzed in a similar

way. A typical line of the csv from 2016 consisted of the following variables: *Trump*, *Clinton*, *Undecided*, *Other*, *poll_id*, *pollster*, *start_date*, *end_date*, *sample_subpopulation*, *sample_size*, *mode*, *partisanship*, and *partisan_affiliation*.

The first three variables are quantitative and are of primary interest for the study. *Trump* contained the percentage of responses for Trump in that poll. *Clinton* contained the percentage of responses for Clinton in that poll. *Undecided* contained the percentage of undecided voters, and *Other* contained the percentage of responses for other candidates. The Other percentage was omitted in select states. The data for 2008 substituted Clinton and Trump with Obama and McCain, respectively. Likewise, the data for 2012 substituted Clinton and Trump with Obama and Romney, respectively. The only data used in the analysis were the data for the major candidates and the data for "Other" (minor) candidates.

The rest of the variables provide potentially useful information about characteristics of the polls. *poll_id* had the id number assigned by Pollster to that poll. The variable *pollster* contained the name of the polling agency, while *start_date* contained the start date of the poll and *end_date* contained the end date of the poll. *sample_subpopulation* indicated whether the poll was of registered voters or likely voters. *sample_size* contained the sample size of the poll and *mode* indicated the method that poll was collected (such as internet, live phone, etc.). The variable *partisanship* indicates if the poll was sponsored by a partisan group, and *partisan_affiliation* indicates the partisan stance of the group that sponsored the poll if it was partisan.

For all elections, polls were only considered if they were conducted after July 1st of the election year. This date was chosen because it represented a point in the election where the two major candidates were usually known and the nomination process was over. For 2016, the conventions occurred shortly after this time marking the beginning of the general campaign. This same date was then used in the analysis of the 2008 and 2012 elections even though those conventions came later. For the 2016 election, the polls were pulled on November 4th for red and blue states and November 5th for the swing states for analysis. The decided deadline for posting predictions was November 5th, 2016 at 11:59 central time. This date was chosen because a prediction made the weekend before an election would contain the vast majority of the pre-election polls, and a weekend prediction allowed more time to comment on the model and analyze the election. The prediction became final at 8:48 central time on November 5th, 2016. For the analysis of 2012, polls were considered if they occurred after July 1st and were released by 12 pm on the Saturday before the election. In 2008, the release dates of the polls were not available. Polls from 2008 were included in the analysis if they were conducted between July 1st and October 31st. However, for 2008 and certain states in 2012 and 2016, Pollster did not have csv files available so they had to be manually created sometimes based on data from FiveThirtyEight. A few states in 2008 and 2012 only had polls from before July 1st, in these cases those polls were included. Other states had no polls available and data from other states had to be used.

2.2 Prior Assumptions

We assume that given a region and a partisan leaning (red states or blue states) we can find a state, preferably the state with the most poll data, to form a prior belief about the other states in that region that also lean the same way. We assume that this prior belief generated by the poll data will at the very least be able to predict the winner of the state and that preferably it will also help indicate the relative competitiveness of that state as well.

We also assume that the outcome in swing states is similar enough to popular vote for national polls to be able to form an accurate prior belief of the outcome of a particular swing state.

It is not expected that these prior beliefs be unbiased and perfectly accurate estimators of the

exact outcome in a state, but that they provide a starting point that can provide a reasonable predictor especially for the states with a small number of polls which are the hardest to predict.

The use of poll data from sources outside the state was chosen because most partisan states did not have polls from a wide variety of agencies. However, there were states that with better polling that were sufficient approximations of the electorate in the state they were being applied to. For swing states, the popular vote was expected to be a good estimation of the votes in swing states. This choice of prior provided a much larger data set. There were over 200 polls used in the prior for swing states for the 2016 election. The number of polls in the prior data was not directly involved in the prediction calculation, but larger data sets do tend to be more stable and accurate than smaller data sets.

To create distinct regions of the country we divide up the country into four distinct groups based on political, demographically, and cultural factors. There is a separate group for swing states. This creates five distinct categories with different priors. The five different priors were: national polls for swing states, Texas for Southern red states, Nebraska for Midwestern red states, California for Western blue states, and New York for Northern blue states. The prior data usually had more polls than the state being predicted. The states used as priors had the following priors: Georgia for Texas, Kansas for Nebraska, Washington for California, and New Jersey for New York. The prior data often had higher-quality polls than other states in that category. Using more than five categories may have worked better, but the limited choices made choosing a prior easier. The states that served as priors were assumed to have more information by election day than the states they would be applied to. There were a few states that were not clear which prior should be used. For red states, the decision on whether a state was in the Midwest or South was primarily based on which state was more similar culturally and demographically, Texas or Nebraska. The addition of another category for the Southeast red states and the Midwest blue states may have worked better, but the small amount of choices in priors made decisions easier. The US Census Bureau divides the country into four main regions (West, Midwest, Northeast, and South). Included below, in Figure 1 and Figure 2, is a map of the US Census Bureau Regions (Figure 1) and a map of the United States color coded to reflect the prior used for that state (Figure 2).

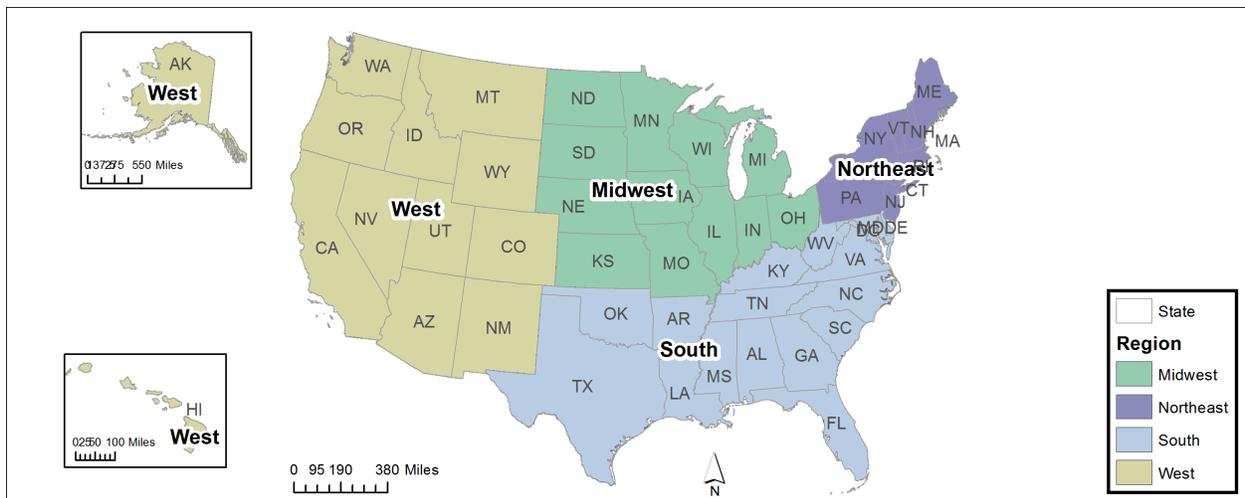
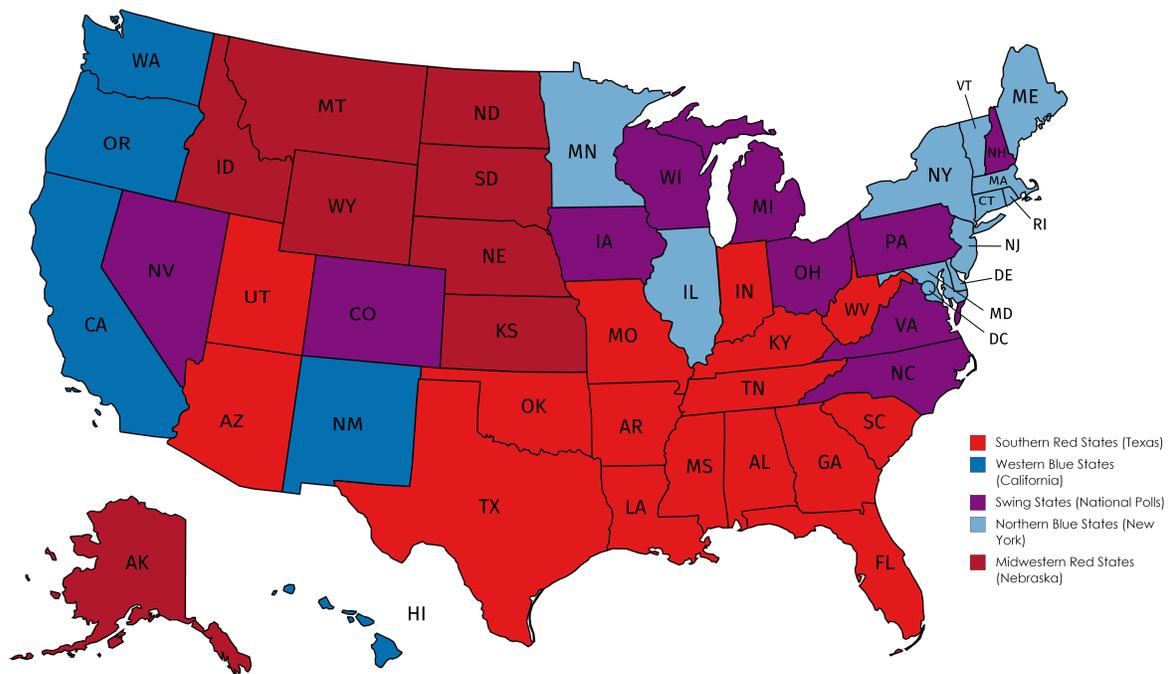


Figure 1: Map of the US Census Regions (taken from census.gov)



Created with mapchart.net ©

Figure 2: Map of the Priors Used for Each State (Created with mapchart.net)

As shown in the figures above, the prior categories are similar to the US Census regions. The main changes are the inclusion of Missouri into the Southern red state category, the inclusion of Utah and Arizona to the Southern red state category, the inclusion of Idaho, Montana, and Wyoming into the Midwestern red state category, and the inclusion of Maryland, the District of Columbia, and Delaware into the Northeastern category. Most of these changes are due to the fact that hypothetical categories for Western red states, and Southern blue states would not be large categories and it would be difficult to find states to use as the prior in those categories.

2.3 Swing States

States were chosen as swing states if the presumed winner was not clear in advance, or if the state did not have a clear leading party that controlled most of the government offices in that state. A few swing states were temporarily considered swing states for a single election cycle if that state did not appear to be voting as usual. The decision of what states are swing states (sometimes called battleground states) is admittedly subjective with different news organizations using different definitions of what constitutes one. However, the organizations typically agree on most of the states that constitute. For the purposes of this project, we chose to use the swing states designated by Politico with two additions: Utah and Arizona. The following thirteen states were considered swing states for at least one election in the analysis: Arizona (2016 only), Colorado (2008, 2012, 2016), Florida (2008, 2012, 2016), Iowa (2008, 2012, 2016), Indiana (2008 only), Ohio (2008, 2012, 2016), Nevada (2008, 2012, 2016) New Hampshire (2008, 2012, 2016), North Carolina (2008, 2012, 2016), Michigan (2008, 2012, 2016), Pennsylvania (2008, 2012, 2016), Utah (2016 Only), Wisconsin (2008,

2012, 2016).

The reasons we included Arizona and Utah as swing states, despite Politico not doing so, are as follows. In 2016 for Arizona, the increased registration of Hispanics and young voters coupled with close polls lead to the decision to consider Arizona as a swing state. Arizona still used Texas as its prior, but Arizona got more attention because of its potential swing state status. In 2016 in Utah, an independent candidate, Evan McMullin got as high as 30% in some polls in October, and it was decided that the strong support for an independent candidate could mean that Trump (the presumed winner) may not win in Utah. This led to the decision to make Utah a swing state in 2016. Although Utah was considered a swing state, it did use Texas as its prior since Utah was not expected to vote like the popular vote. As a final note, for Indiana in 2008, President Obama got a surprising amount of support considering the strong red voting history of that state.

2.4 Model Formation

A series of simplifying assumptions were made about the election to make the analysis easier. First, it was assumed that, over the campaign, people who reported themselves as decided would not change their minds, and therefore, all polls within the timeline of analysis could be treated equally.

We assume that the poll data of all states is normally distributed. We also assume that data from another similar state is a accurate predictor of the polls in another states. The number of polls was small in some cases, but the amount of people polled in a single poll is usually at least several hundred votes with some polls having a thousand or more responses. This is a simplifying assumption that may not actually be true and needs to be examined in further research.

This study only focused on the prediction of major candidates. If a candidate gets 5% of the vote nationally their party is considered a major party. A minor candidate is a candidate that did not get 5% of the national vote on election day. We assume that the only major candidates in 2008 and 2012 would be the Republican and Democratic candidates, and that in 2016 we assume that Gary Johnson would also be a major candidate in addition to the Republican and Democratic candidates. More details on how Gary Johnson's support was calculated and comments on why he was included can be found in the subsection titled "The Case of Gary Johnson".

It was decided to use the Maximum Likelihood Estimators (MLE) for μ and σ^2 , and to not use the binomial approximation to the normal distribution. The maximum likelihood estimator for σ^2 was chosen because it was a built-in function in the programming language being used which saved programming time. The results were then after all calculations so that the total of the probabilities equaled 1. This approach would also be far easier to compute and implement, making this the best option at the time. The level of undecided voters in the polls included in the analysis did not change much over time, but there were slightly more undecided voters in the early polls. Normalizing before the analysis would require more calculations and computation time, and the prediction for 2016 had to be conducted in less than 36 hours so that the final round of polling could be included in the analysis. A change could have been made in the analysis for 2008 and 2012 to include methods like normalization before the analysis. This change was not done because the goal was to test the performance of the model under similar circumstances.

As discussed previously, this model utilizes a Bayesian approach using the poll data of other similar states as priors. The model predicted the final percentages of the votes within each state using the maximum likelihood estimators for the mean and variance for the polls from the prior state, and the mean and variance of the polls from the state being predicted. The description of the Gaussian conjugate prior and its derivation are based on the derivation in [6]. Intermediate steps are omitted. We assume the distribution of every state is normal. We also assume the distribution of

national polls is normal. We assume that the variance is fixed, and can be sufficiently approximated by the MLE. The prediction calculation is defined below.

The Maximum Likelihood Estimators $\hat{\mu}$ and $\hat{\sigma}^2$ are given as follows.

Maximum Likelihood Estimators for σ^2 and μ

$$\hat{\mu} = \frac{1}{n} \left(\sum_{i=1}^n p_i \right) \quad (1)$$

$$\hat{\sigma}^2 = \frac{1}{n} \left(\sum_{i=1}^n (p_i - \hat{\mu})^2 \right) \quad (2)$$

Here, p_i is the result of a poll for a given state and \bar{p} is the mean of those results in that state.

Gaussian Conjugate Prior Calculation Given the sample mean \bar{p}_{c_0} and variance $\hat{\sigma}_{c_0}^2$ 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 sample variance $\hat{\sigma}_c^2$ of the polls from the state being analyzed for a candidate c. Then given a set D of n polls: $\{D = p_1, p_2, \dots, p_n\}$. Using Bayes Rule, the likelihood of observing this sequence of polls, given μ_c and σ^2 , is:

$$p(D|\mu_c) \propto \exp\left(-\frac{n}{2\sigma_c^2}(\bar{p}_c - \mu_c)^2\right) \propto \mathcal{N}(\bar{p}_c | \mu_c, \frac{\sigma_c^2}{n}) \quad (3)$$

Hence, the natural conjugate prior is of the form,

$$p(\mu_c) \propto \exp\left(-\frac{n}{2\sigma_{c_0}^2}(\mu_c - \mu_{c_0})^2\right) \propto \mathcal{N}(\mu_c | \mu_{c_0}, \sigma_{c_0}^2). \quad (4)$$

Equations (1) and (2) imply:

$$p(\mu_c|D) \propto p(D|\mu_c, \sigma_c^2)p(\mu_c|\mu_{c_0}, \sigma_{c_0}^2) \propto \exp\left[-\frac{n}{2\sigma_{c_0}^2} \left(\sum_{i=1}^n (p_i - \mu_c)^2\right)\right] \times \exp\left[-\frac{n}{2\sigma_{c_0}^2}(\mu_c - \mu_{c_0})^2\right] \quad (5)$$

using Bayes Theorem, which states

$$P(\mu_c|D) = \frac{P(D|\mu_c)P(\mu_c)}{P(D)}. \quad (6)$$

Based on the derivations in equations (3),(4), and (5), the posterior distribution of support for a candidate c is a Normal distribution with updated mean and variance parameters, which can be calculated as follows:

$$\mu_{post} = \frac{\sigma_c^2}{n\sigma_{c_0}^2 + \sigma_c^2} \bar{p}_{c_0} + \frac{n\sigma_{c_0}^2}{n\sigma_{c_0}^2 + \sigma_c^2} \bar{p}, \quad \sigma_{post}^2 = \left(\frac{1}{\sigma_{c_0}^2} + \frac{n}{\sigma_c^2}\right)^{-1} \quad (7)$$

Since the true mean and variance of the prior and data are unknown in our final calculation we must use estimators. Using equations (1) and (2) for the MLEs, we obtain estimators for the posterior means and variances of the following form:

$$\mu_{post} = \frac{\hat{\sigma}_c^2}{n\hat{\sigma}_{c_0}^2 + \hat{\sigma}_c^2} \bar{p}_{c_0} + \frac{n\hat{\sigma}_{c_0}^2}{n\hat{\sigma}_{c_0}^2 + \hat{\sigma}_c^2} \bar{p}, \quad \sigma_{post}^2 = \left(\frac{1}{\hat{\sigma}_{c_0}^2} + \frac{n}{\hat{\sigma}_c^2}\right)^{-1} \quad (8)$$

2.5 Normalization of Predicted Values

The Gaussian conjugate prior was used to calculate the support for an individual candidate in a specific state for a specific election. The calculation was made for each candidate and then the results were normalized so that the proportions summed to 1 with the following formulas:

Normalization Calculation

$$r_{p_i} = \frac{r_{ppst}}{r_{post} + d_{post} + l_{post} + o_{post}} \quad (9)$$

$$d_{p_i} = \frac{d_{post}}{r_{post} + d_{post} + l_{post} + o_{post}} \quad (10)$$

$$l_{p_i} = \frac{l_{post}}{r_{post} + d_{post} + l_{post} + o_{post}} \quad (11)$$

$$o_{p_i} = \frac{o_{post}}{r_{post} + d_{post} + l_{post} + o_{post}} \quad (12)$$

Where r_{p_i} , d_{p_i} , l_{p_i} , and o_{p_i} are the normalized final prediction values for the republican, democratic, libertarian (for 2016 only), and minor candidates respectively, and r_{post} , d_{post} , l_{post} , and o_{post} are the original predictions for the republican, democratic, libertarian (for 2016 only), and minor candidates respectively. If the libertarian and minor candidates were not included in the analysis, the respective terms were omitted from the calculation.

2.6 Sample Prediction

Included below, in Table 1 and Table 2, are the inputs and outputs of the model for the states of Colorado and Wyoming in 2016. The table expresses the variability in standard deviations instead of the variance because standard deviation has the same units as the mean. Only the posterior column is standardized for the sums to equal one. Since the polls are treated as independent and identically distributed (i.i.d) random variables, the posterior standard deviation is of the form $\frac{\sigma}{\sqrt{n}}$. This is why Colorado, which had 36 polls, has a much smaller standard deviation than Wyoming which had 4 polls. This caused a great underestimation of the variability of the predictions made by the model.

Table 1: Sample Prediction for Colorado in 2016

Candidate	Prior	Data	Posterior
Clinton Mean	0.431	0.444	0.471
Clinton SD	0.03	0.027	0.005
Trump Mean	0.397	0.391	0.415
Trump SD	0.029	0.035	0.006
Other Mean	0.039	0.059	0.059
Other SD	0.017	0.052	0.008
Johnson Mean	0.073	0.048	0.054
Johnson SD	0.023	0.046	0.008
Number of Polls	219	36	

Table 2: Sample Prediction for Wyoming in 2016

Candidate	Prior	Data	Posterior
Clinton Mean	0.346	0.253	0.315
Clinton SD	0.025	0.046	0.017
Trump Mean	0.511	0.608	0.624
Trump SD	0.03	0.053	0.02
Other Mean	0	0	0.001
Other SD	0.001	0.069	0
Johnson Mean	0.041	N/A	0.06
Johnson SD	0.029	N/A	0.029
Number of Polls	7	4	

As shown in the Tables above, the posterior gets closer in distribution to the data as the number of observations (polls) increases. Since the polls were treated as independent identically distributed random variables, the standard deviation decreases with each new observation.

2.7 Explanation of Assumptions

The Gaussian conjugate prior resulted in simpler calculations than other choices of priors would have. As used in this model, we assume polls are identically and independently distributed to avoid requiring the estimation of the covariances between polls. However, polls are likely not independent, but the covariance between polls may have a minimal affect on the calculation on the variance of the prior distributions. Additionally, while polls have small biases that cause them to have slightly different distributions when the polls are pooled together, much of this bias should cancel out. The number of polls in the state varied widely, but was typically below the standard requirement of 30 polls. It was also unknown how long it would take to write the programs to run the model and how long it would take to make a prediction. The programs needed to be prepared in a few months to make a 2016 prediction, and the programs had to be fast enough to make the predictions in a couple of hours so that all the poll data could be used. Although the Gaussian conjugate prior is not the ideal model, but since it easy to program and easy to run which made it the best option in this case. Future research needs to examine how to make models that are both theoretically sound and computationally efficient and begin well in advance of the election they are attempting to predict.

Using different states as the prior does introduce some small bias in the prediction. Since the normal conjugate prior approaches the distribution of the data as the number of observations increases, the prior did not need to be an exact approximation if the state being analyzed had sufficient poll data, like in the case of Texas. Some of the larger states like Georgia and Illinois have almost as much poll data as the state used as the prior, and the posterior distribution is similar to the distribution of poll data from that state. As seen in Table 1 and Table 2, the prior distribution had a stronger influence in Wyoming which had 4 polls than in Colorado which had 36 polls. But the bias introduced by the prior tends to be consistent and could perhaps be easily adjusted by shifting the prior to better reflect the state being predicted.

2.8 The Case of Johnson in 2016

If a candidate gets 5% of the vote nationally their party is considered a major party. A minor candidate is a candidate that did not get 5% of the national vote on election day. The inconsistency

in the inclusion of other candidates in polls from all election years made estimating support for other candidates difficult. For 2008 and 2012, an analysis of minor candidates was not included because the combined votes for third party candidates were less than five percent. Unlike 2016, there was not a consistent pattern of polls above 5% for any third party or independent candidate. Like 2016, the data for 2008 and 2012 did not have consistent information on the support of minor candidates. If an analysis on the minor candidates was conducted, it probably would not have been more accurate because most polls had no data on the minor candidates meaning the mean of other polls would be around 0. Not predicting the support of minor candidates may not be ideal, but it was practical.

It was also assumed based on the early poll data in September that Gary Johnson would get at least five percent of the vote in 2016, and therefore should have a predicted mean for all states. Because of this assumption, a method to approximate his support was necessary. However, the poll data on Johnson was limited and a measurement called the “Other-Johnson Factor” was created. The Other-Johnson Factor measured the percentage support of non-major party candidates (not Clinton nor Trump) that was for Johnson. If the prior state had data on Johnson (i.e. California) then the Other-Johnson Factor was based on the prior state.

Other-Johnson Factor

Let J be the average poll support for Johnson, and O be the average poll support for other candidates besides Trump, Clinton, and Johnson. Then

$$\text{Other-Johnson Factor} = \frac{J}{J + O} \tag{13}$$

In most cases, the Other-Johnson Factor was based on national polls. For the swing states, only national polls with Johnson data were included. In October, Evan McMullin, an independent candidate, appeared in some polls in Utah at around 20% to 30%. At this point, it was decided that Evan McMullin should be included in the analysis of Utah using the same method as the Other-Johnson Factor. Evan McMullin did not have enough support in any other state to be considered as a major candidate in elsewhere. In the case of the Western blue states, where the Other-Johnson Factor was based on the poll data of California, the presence of polls without minor candidates lowered the average significantly and caused an underestimation of Johnson in the Western blue states. In Nebraska, there was a very limited amount of information on other candidates, and this made synthesis of other support difficult in the Midwest. In 2016, if other support was not specified it was calculated by subtracting the sum of Trump's support, Clinton's support, and undecided voters from 1. A poll respondent can only support Trump, Clinton, or another candidate, or be undecided, making the total probability of these numbers 1.

3 Results

3.1 Challenges of the 2016 Election

The 2016 election proved to be difficult to predict. As discussed in [11], a common theme in the media message and the models was that Clinton was likely to win the election. Florida, Michigan, North Carolina, Pennsylvania, and Wisconsin were frequently missed in the predictions of multiple models. The tested model called 88% of the states, missing Florida, Ohio, Michigan, North Carolina, Pennsylvania, and Wisconsin. Since all polls were weighted equally regardless of the date of the poll, the model did not adapt well to trends over time. As Election Day approached, Trump began to have a small lead in the polls in Ohio. But since the tested model weighted all polls equally, the polls with Trump leads did not raise his average enough for the model to

predict a Trump win. The five frequently missed states were close with margins of less than 1% in Pennsylvania, Michigan, Florida, and Wisconsin. The exact reasons for the errors in the polls and models are unknown at this time. Polling and election prediction models should be examined in light of this election. It is possible that the 2016 election was an outlier that may have been unpredictable. As discussed previously the model was applied to the 2008 and 2012 elections, using poll data with the same poll inclusion criteria as the 2016 model.

3.2 Accuracy of Models to Predict the Winners of States

Below, in Table 3, we compare the accuracy the various models (including the tested model) of predicting the winning candidate in each of the fifty states plus Washington DC (if available).

Table 3: Percentages of Winners Called Correctly

Race	Tested Model	RCP ¹	PEC ²	538 ³	PW ⁴
2008 Accuracy	0.98039	0.96078	0.98039	0.98039	N/A
2012 Accuracy	1	0.98039	0.98039	1	0.98039
2016 Accuracy	0.88235	0.92157	0.90196	0.90196	0.90196
Average Accuracy	0.95425	0.95425	0.95425	0.96078	0.94118

¹ RealClearPolitics, [8],[9],[10]

² Princeton Election Consortium: election.princeton.edu/

³ FiveThirtyEight Polls Plus Model: fivethirtyeight.com

⁴ Predict Wise: predictwise.com/

3.3 Other Methods of Accuracy Measurement

All models achieve highly similar accuracy in predicting the winners in each state. In 2008 and 2012, the tested model had the same predictions of the winning candidates as FiveThirtyEight. Measurements such as Brier scores or root mean square error help to provide more clarity in the true accuracy of predictions by measuring the accuracy of the probabilistic and mean predictions. The tested model predicted the distribution of votes among the major candidates. To compare the accuracy of mean prediction the root mean square errors (RMSE) of both all states and swing states, were calculated for 2008, 2012 and 2016 for the tested model, FiveThirtyEight's model and the RealClearPolitics Average ([8],[9],[10]). In all years, RealClearPolitics ([8],[9],[10]) did not have a complete data set on the states so only the swing states were analyzed. The RealClearPolitics Average ([8],[9],[10]) did not always sum to 1, and the numbers were normalized before the root mean square error was calculated. States were weighted equally, and Washington D.C. was treated like a state for the purpose of this calculation.

3.4 An Analysis of the Root Mean Square Error of the Model

The root mean square error helps provide a comparison the accuracy of a model's prediction of the final vote percentages. If the model normalizes probabilities to 1, the sum of overestimations equals the sum of underestimations. This is not the only way to measure the accuracy of mean

prediction. For example, the root mean square error can be calculated for the relative margins between candidates. This method was chosen because it can be used to provide an estimate of the margin of error of these models. The root mean square error was calculated with the following formula:

Root Mean Square Error Calculation

$$RMSE_{model} = \sqrt{\frac{\sum_{i=1}^{n_{states}} \left(\frac{|r_{a_i} - r_{p_i}| + |d_{a_i} - d_{p_i}| + |l_{a_i} - l_{p_i}| + |o_{a_i} - o_{p_i}|}{2} \right)^2}{n_{states}}} \quad (14)$$

where r_{a_i} , d_{a_i} , l_{a_i} , o_{a_i} are the actual votes for the Republican, Democrat, Libertarian (2016 only), and other candidates in state i , respectively, and r_{p_i} , d_{p_i} , l_{p_i} , o_{p_i} are the predicted votes for the Republican, Democrat, Libertarian (2016 only), and other candidates in state i , respectively. Note that the terms l_{a_i} & l_{p_i} are not used in 2008 and 2012 where the Libertarian party was not considered to be a major party by predictors. For two-party predictions only the terms r_{a_i} & r_{p_i} & d_{a_i} & d_{p_i} were used in the calculation. We divide by 2 in the numerator because every underestimation of one candidate results in an overestimation of one candidate since all the proportions sum to one.

Referring to the sample prediction from above: the predicted outcome of the tested model for Colorado in 2016 was 47.1% for Hillary Clinton, 41.5% for Donald Trump, 5.5% for Gary Johnson and 5.9% for other candidates. The actual voting result was 48.2% for Hillary Clinton, 43.3% for Donald Trump, 5.2% for Gary Johnson and 3.3% for minor candidates. This would give the prediction for Colorado in 2016 a root mean square error of 2.9. The root mean square error was calculated for each individual state for the three models being compared. Then, the root mean square error of all of the predictions by the particular model were calculated for each election year. Since predicting the exact voting results is more important in swing states where the winner is less predictable, an analysis on the accuracy of swing state prediction was also conducted.

An estimate of a prediction for only the two major candidates was made by taking the estimates for the Republican and Democrat, and then proportionally normalizing that support to equal one by using formulas 8 through 11. Since the vast majority of votes goes to either of the two major parties, predicting the performance of those candidates relative to one another is important. All of these errors are large enough for it to be reasonable to expect models to not always call the winner of the election.

Below, in Table 4, we compare the Root Mean Square Error of the tested model, RealClearPolitics model ([8],[9],[10]), and the FiveThirtyEight Polls Plus model in both all states and in swing states. We use the list of swing states used by the model to calculate the swing state error. Utah and Arizona did not actually have a close election in 2016, and their RMSE appeared to be outliers, so we report the swing state average without Arizona (AZ) and Utah (UT) considered as swing states. As stated previously only the root mean square error for RealClearPolitics ([8],[9],[10]) in swing states was calculated because not all states had a RealClearPolitics average ([8],[9],[10]). The table includes the root mean square error for all candidates and the root mean square error for the two major candidates. The table also includes the root mean square error for swing states. The Two-Party root mean square error is based on the two-party estimates discussed in the previous paragraph. The all candidate average row contains the average of a model's prediction for all candidates in the 2008, 2012, and 2016 election. The 2-Party Average row contains the average of a model's root mean square error for predicting 2-Party support. The 2-Party Average Compared to 538 row contains the percent accuracy of the model compared to 538 for the appropriate year and category. The 2-Party Average compared to RealClearPolitics ([8],[9],[10]) row contains the accuracy of the tested model root mean square error for swing state compared to the RealClear-

Politics Average ([8],[9],[10]) root mean square error for swing states. Note the root mean square error below are in percentages such that a root mean square error of 1 represents a error of 1%.

Table 4: Root Mean Square Error for Various Models

Race	TM ¹	TM SS ²	RCP ³ SS ²	538 ⁴	538 SS ²
2008 All Candidates	3.54740	3.14788	4.23389	3.19332	1.66958
2008 2-Party	2.89669	2.57051	3.63513	3.03050	1.47846
2012 All Candidates	3.25139	1.94492	2.33511	2.38019	1.29790
2012 2-Party	2.37053	1.17163	1.61076	1.98642	0.93420
2016 All Candidates	6.82013	6.42335	8.23311	5.37952	4.14228
2016 2-Party	3.95985	3.03986	1.89412	3.81296	2.41263
2016 SS ² w/o ⁵ UT and AZ		3.99534	3.32952		3.56511
2016 SS ² w/o ⁵ UT and AZ 2-Party		3.14325	2.04295		2.31948
All Candidate Average	4.53964	3.83872	4.93404	3.65101	2.36992
2-Party Average	3.07569	2.26067	2.38	2.94329	1.60843
2-Party Average Compared to 538 ⁴	.95695	0.71148	.67581		
2-Party Average Compared to RCP ³		1.05279			

¹ TM: Tested Model

² SS: Swing State

³ RealClearPolitics [8],[9],[10]

⁴ FiveThirtyEight Polls Plus Model: fivethirtyeight.com

⁵ w/o : without

As shown in Table 4, the tested model was 95.329% as accurate as the FiveThirtyEight model in all states, but was only 68.727% as accurate in swing states. The model was 5.859% more accurate than the RealClearPolitics Average ([8],[9],[10]). Considering the relative simplicity of the model, the model performed very well overall. However, the model did not work well in swing states.

3.5 Interpretation of the RMSE

A similar Bayesian model could not be found, so it was unknown how this model would perform. A fundamental model, which is a regression model using political and economic data, particularly a model with separate regressions for partisan and swing states may perform better. However, the model performed very well in other states, and shows promise as a method for determining relative support in partisan states. Further research is needed to see if poll data from other states is a better estimator for voter behavior in partisan states than fundamental models.

3.6 Selected Predictions and Results for Selected States for 2016

Included below is a table of selected two-party predictions for 2016 with actual two-party results and two-party root mean square error for individual states. A two-party analysis better showcases the accuracy of a model, because it ignores minor candidates, which are both difficult to predict and are only a small portion of the vote. The table includes all swing states, the red and blue state with the highest error (West Virginia & Hawaii), the red and blue state with the lowest error (Texas & Massachusetts), and two close partisan states (Arizona & Minnesota). Wyoming is also included since it was used as an example prediction. Trump was usually underestimated

by the model because polls regularly underestimated President Trump's support. This consistent underestimation of President Trump's is probably the cause of the higher modeling error in 2016. Below, in Table 5, the two party predictions, results, and root mean square error are displayed for selected states.

Table 5: Selected Two Party Predictions, Results, and Root Mean Square Errors for 2016

State	Trump Actual	Trump Predicted	Clinton Actual	Clinton Predicted	2 Party RMSE
West Virginia	0.721	0.62	0.279	0.38	10.1
Texas	0.547	0.54	0.453	0.46	0.7
Arizona	0.519	0.508	0.481	0.492	1.1
Hawaii	0.312	0.354	0.688	0.646	4.2
Massachusetts	0.353	0.352	0.647	0.648	0.1
Minnesota	0.492	0.449	0.508	0.551	4.3
Wyoming	0.757	0.664	0.243	0.336	9.3
Colorado	0.473	0.468	0.527	0.532	0.5
Florida	0.506	0.485	0.494	0.515	2.1
Iowa	0.551	0.507	0.449	0.493	4.4
Michigan	0.501	0.461	0.499	0.539	4
Nevada	0.487	0.492	0.513	0.508	0.5
New Hampshire	0.498	0.468	0.502	0.532	3
North Carolina	0.519	0.489	0.481	0.511	3
Ohio	0.543	0.497	0.457	0.503	4.6
Pennsylvania	0.504	0.466	0.496	0.534	3.8
Virginia	0.471	0.455	0.529	0.545	1.6
Wisconsin	0.504	0.469	0.496	0.531	3.5

3.7 Accuracy of the Model in Swing States

The choice of national polls as the prior for swing states may have affected the accuracy in the swing states. While the model had good performance overall compared to other models, the model's performance in swing states was relatively poor. The use of national polls to represent swing states felt natural. It made more sense to use national polls for the swing states since they vote relatively close to the popular vote than any other state. It was known that national polls would not be a perfect fit. After a more thorough examination of the assumption that swing states vote like the national vote, it was concluded that national polls should not be used as the prior. Swing states are the most important states to predict, and the method failed in the area that mattered the most.

3.8 Accuracy of the Predictions for Minor Candidates

Predicting the support of third-party and independent candidates was difficult and was often underestimated by the model. National polls may include the minor candidates in their questions about support, but the state polls often did not ask about the minor candidates. The support for minor candidates is often small, and in most states the winner is clear. This lack of information on minor candidates did not affect the model's ability to call a winner, but most of the errors in prediction were related to an underestimation of minor candidates. Since some polls do not include minor candidates this lowered the average of the support of those candidates, which sometimes caused an

underestimation of minor candidates. In some states the only candidates on the ballot were Trump, Clinton, and Johnson so this underestimation did not matter much, but in other states, particularly those where Jill Stein (the Green Party candidate) was on the ballot, this underestimation lowered the accuracy. It may be better to ignore minor candidates and instead predict the levels of support in a two-party race. A two candidate prediction approach appears to be the most accurate way to predict and compare the support of the two major candidates.

In the 2016 election, the prediction of votes for Gary Johnson were especially problematic. Johnson got under five percent of votes nationally, and Johnson was only included in the analysis because of an assumption he would get at least five percent of the national vote on Election Day. The “Other-Johnson Factor” did not work very well because in some national polls Johnson was the only minor candidate included. This meant that in the final prediction the “Other-Johnson Factor” was over .9999, thus implying Johnson would get 99.99% of all the votes for the minor candidates. However, Johnson did not receive 99.99% of the votes for minor candidates and the other minor candidates were underestimated.

3.9 Accuracy of the Model in Limited Information States

In states with limited poll information, the errors were larger, but these states often had a clear leader so polls were not needed to be conducted as often as in closer states. For example, in West Virginia, the Republican candidate had over a five percent lead in all three elections studied. West Virginia is known to be a red state, but the tested model’s prediction of West Virginia in 2012 was seven percent off. However, the model still called the winner so this error did not affect the accuracy of predicting the Electoral College. But in swing states where the winner may be decided by a percentage within the past root mean square error of the means for swing states, accuracy matters more. The inaccuracies in predicting the winner of the 2016 election mattered more than in 2012, where the swing states were not as closely decided. Only one state in 2012 was decided by less than one percent, but in 2016 this happened in four states. Given the unusually large size of the errors in poll-based models for the 2016 election, it is possible that the poll data itself could have been bad. The 2016 election may have been unpredictable given the poll data, but that does not mean the models can not be improved. The FiveThirtyEight model did the best at predicting means probably because it used methods to address non-response and polling agency bias.

4 Discussion

It is important to interpret this study in context. As an undergraduate research project, there were knowledge and time limitations with this model. This project was completed in approximately four months so that a prediction of the 2016 election could be made. The model needed to predict the outcome for all 50 states and Washington D.C. quickly to allow time for the predictions and an commentary on the predictions to be posted before the election. The model and computer programs to run the program were not finished until a few weeks until the election. The simplifications and assumptions made the study possible. The model may be relatively simplistic compared to models discussed in [1], [2], [4], and [5], but it performed well at predicting a two-party race. Polls are not completely independent in practice and opinions change over the campaign. This version of Gaussian conjugate prior assumes that the observations are independent random variables. The Gaussian conjugate worked quite well in predicting the winner, but it greatly underestimated the variability of polls. While the tested method was not designed to primarily create a probabilistic representation of winners in the states, the credible intervals were narrow and frequently failed. Looking forward to future elections, the primary author plans to try more complicated models that

are currently unfeasible for her at this time. A possible model would be a binomial multilevel time series Bayesian model, with adjustments for poll bias to predict two-party support. The end goal of this line of research is a time-series multilevel Bayesian model with post-stratification using poll data that can be implemented a created to predict an upcoming election.

4.1 Challenges of American Election Prediction

Some of the challenges faced building this model were how to define a swing state and how to break the US into regions. The decided definitions for both the regions of the US and swing states are admittedly somewhat subjective. Great care was taken to make the decisions as informed as possible, including pouring over past election results to look for trends of states that voted similarly to each other, and examining if states were competitive in the past. However, one of the benefits of a Bayesian analysis is that it allows the modeler to include prior beliefs in the model, even if they are subjective.

The opinions of the American electorate are constantly changing and the 2016 election exposed the limitation of models that worked for previous elections. By the Central Limit Theorem, as the number of polls approaches infinity the distribution normalizes. This normalization may not be happening fast enough in the context of a presidential campaign. The number of polls is small in some cases meaning that other distributions may work better for poll analysis. The binomial approximation to the normal distribution would likely be more appropriate in this situation. Other distributions and their respective conjugate priors may work better than the Gaussian conjugate prior.

The limitation of poll-based election prediction is that there is currently a low number of elections with easily accessible poll data. There have been promising results with a method that uses a fundamental model to predict the winner of states, as discussed by [4] and used by PredictWise. The fundamental approach has produced similar results to poll-based methods for the past two elections, but it is difficult to tell if a fundamental based model is truly a better method of predicting state-level support. The definition of accuracy for Presidential election prediction is difficult to distinctly define because accuracy can be determined by Brier scores, root mean square error, or correct prediction of the Electoral College. As a result, it is difficult to distinguish which models are more accurate than others.

It is important that predictors understand that, while they may be able to predict the winner, estimates of the mean and variability are usually inaccurate in simplistic models. While the errors in predicting voting results always existed, 2016 was the only year where the errors caused predictors to inaccurately predict the winner of the election. The problem was not necessarily that the error in predicting means was significantly larger, but rather that the error affected the prediction of the winner.

The major drawback of the FiveThirtyEight model is that it is proprietary, and therefore, not available to the public. While it seems to be more statistically sound compared to other models, its specific methodology is unknown. The Princeton Election Consortium model is open source, but it is based on assumptions that seem not to hold. The lack of multiple good open-source models for presidential elections makes the academic study of polling analysis by statisticians important work. There are numerous studies based on a “fundamental” model using data from other sources, such as the economy, as discussed in [2] & [4]. Fundamental models have been successful at predicting the winner of the popular vote. Most of the time, the winner of the popular vote wins the Electoral College, but there are cases like 2000 or 2016 where the winner of the popular vote does not win the presidency. Polls are probably improving in both their methods and volume. Even across the elections analyzed in this study the number of polls began to increase over time. The difficulties in

creating a representative poll means the poll data are subject to non-response bias. Polls may have flaws, but the integration of multiple approaches to address possible biases can improve accuracy.

4.2 Future Research

While the results of the model are promising, it has numerous issues that need to be addressed in future research. Since this model was originally designed to predict the 2016 election, the methodology did not always consider how to extend the model to other elections. The computation time of the model was a major concern that influenced model design, but the prediction process was much faster than anticipated and normalization before the prediction calculation could have been implemented had we known that it would not adversely affect the ability for the model to be used to quickly predict an election.

A more formalized definition of a swing state is needed since the swing state definition used in this model based on the 2016 election did not translate well to 2008 and 2012. Specifically, the consideration of Utah and Arizona as swing states proved problematic because Arizona was not competitive and Evan McMullin in Utah did not capture enough of the vote to affect the outcome. In future research, there should be predefined definitions of swing states to reduce bias and make decisions better informed by data rather than opinion. The future priors should be more similar to the US Census regions and should divide up states into regions and have a prior for both the red and blue states in that region. A possible more defined method to decide swing states would be to define a swing state as a state that was won by both a Democratic candidate and a Republican candidate in the past four presidential elections.

Had it been known that Johnson's support would decrease as the campaign approached and that Johnson would begin to disappear from the state-level data, Johnson would have not been included in the analysis. Future research should consider ignoring all candidates not from the two major parties or develop a better method to synthesize third party and independent candidate support. Likewise, Evan McMullin probably should have been ignored because it was very difficult to estimate his support and he did not attract enough support to jeopardize President Trump's chance of winning Utah.

The Beta-Binomial conjugate prior could also be used. A new set of models with some of these changes is currently being implemented as a new research project for the author. Future undergraduate projects can implement similar processes in future elections using these proposed changes.

A possible application of this method would be to Senate and House elections. Senate and House elections may not have a lot of polling information. The number of polls varies widely based on the particular race. Some areas get more polls than others, and an approach like the one this model used could help improve the prediction of Senate and House races. Priors would have to be chosen to reflect multiple factors, including incumbency advantage, demographics, and the political lean of that area.

The limitations of poll-based election prediction is that there is currently a low number of elections with easily accessible poll data. There have been promising results with a method that uses a fundamental model to predict the winner of states, as discussed by [4] and used by PredictWise. The fundamental approach has produced similar results to poll-based methods for the past two elections, but it is difficult to tell if a fundamental based model is truly a better method of predicting state-level support. The definition of accuracy for Presidential election prediction is difficult to distinctly define because accuracy can be determined by Brier scores, root mean square error, or correct prediction of the Electoral College. It is difficult to distinguish which models are more accurate than others. It does appear, though that the RealClearPolitics average ([8],[9],[10]) is not

a good method to predict exact voting results.

4.3 Political Bias Disclosure

It is important to mention the primary author's political biases could have subtly affected the results of the study. The primary author is a moderate Republican that voted for an unregistered write-in candidate in the general election and for Marco Rubio in the primary. The goal was to remain as objective as possible, but personal biases may have played a minor role in model design and analysis. It is important to note that nearly all models could be subject to influence by personal biases of the designer(s) of the model. It is unlikely that bias alone is responsible for all the errors in models, but it could be possible that bias affected the model and interpretation of the data.

4.4 Conclusion

The nature of American Presidential elections is that they only happen once every four years. This makes perfecting a model difficult because, while a model can be tested retrospectively at any time, it may not be able to be applied to a future election for as many as four years. The answer to predicting elections and estimating public opinion may not be simple, but is a worthy topic to study. Every election is different. The changes in both the candidates and the state of the nation mean that older data and older methods may not work in the future. If methods to more accurately predict voting results could be found they could be possibly applied to nomination processes and some foreign elections, which are hard to accurately predict because of the higher level of precision needed to predict a winner given the possibly larger number of candidates. An application to estimation of public opinions as suggested in [7] may exist. The prediction of elections may not be trivial and the methods discovered could solve other problems in political science and beyond. In light of the 2016 election, political science statisticians must begin to examine the assumptions and distributions used in models and try to determine a more accurate method to predict elections. If a solid approach is discovered, it may not work indefinitely and could have to be reexamined for every unique election cycle. But presidential election prediction is an important field to study given its potential for both statistical outreach and applications to other problems.

5 References

- [1] W.F. Christensen and L.W. Florence, Predicting Presidential and Other Multistage Election Outcomes Using State-Level Pre-Election Polls. *Amer. Stat.*, 62:1(2008), pp. 1-10
- [2] A. Gelman and G. King, Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?, *British Journal of Political Science* 23:04(1993), pg. 409-451 [3] Huffington Post, Pollster, <https://elections.huffingtonpost.com/pollster>,(November 3, 2016)
- [4] P. Hummel and D. Rothschild, Fundamental Models for Forecasting Elections at the State level, *Electoral Studies*, 35(2014), pp. 123-139
- [5] K. Lock and A. Gelman, Bayesian Combination of State Polls and Election Forecasts. *Political Analysis*, 18(3) (2010), 337-348.
- [6] K. Murphy, Conjugate Bayesian analysis of the Gaussian distribution (Tech.). <https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf> (2007, October 03)
- [7] D. Park and A. Gelman, and J. Bafumi, Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls, *Political Analysis*, 12(2004), pp. 375,385
- [8] RealClearPolitics, Battle for White House, https://www.realclearpolitics.com/epolls/2012/president/2012_elections_electoral_college_map.html (November 6, 2012)
- [9] RealClearPolitics, Battle for White House, https://www.realclearpolitics.com/epolls/2016/president/2016_elections_electoral_college_map.html (November 5, 2016)
- [10] RealClearPolitics, RealClearPolitics Electoral College, http://dyn.realclearpolitics.com/epolls/maps/obama_vs_mccain/index.php (November 4, 2008)
- [11] N. Silver, The Real Story Of 2016, <https://fivethirtyeight.com/features/the-real-story-of-2016/> (January 19, 2017)