Case Study 3: Election Prediction

Bob Ding, Becca Erenbaum, Grace O'Leary, Rena Zhong

11/1/2020

1. Introduction

The outcome of the 2016 election not only stunned the nation, but also sent shockwaves through the statistical and polling communities. Over the course of the election year, in 2016, up until the week of the election, poll predictions of Hillary Clinton's likelihood of beating Donald Trump ranged from 71%- 99% probability [1]. So, when Trump beat these odds, the polling industry lost a lot of trust from the general public [2]. This small, specialized industry that is the political polling business, has a great deal of influence on how the election is portrayed in the news media, voter decisions, and candidate partisan policy initiatives. [1]. Million dollar decisions on advertising and campaign strategy are dictated by polls. Polls conducted early in the election year are a weak predictor of election outcomes because the general public has paid less attention to the race and is less knowledgeable of the candidates' platforms. Voters that tend to sway between parties often report that they are undecided. However, these are arguably the most important opinions in predicting the election outcomes. [3] As the election nears, polls become more accurate. That is where historical prediction models have been lacking – they fail to take into account the differential accuracy in poll results. Historical models, regression based models that rely on outcomes from past elections and structural factors, predict election outcomes at a single point in time which renders high levels of uncertainty [4]. Drew A. Linzer developed a dynamic bayesian forecasting model to predict the U.S. presidential election at the national and state level that combines these historical models with everchanging poll updates. Linzer's model uses hierarchical specification to handle states being polled on different days and takes into account sampling errors of the polls and national campaign effects [4].

In this report we aim to answer the following 4 questions:

- 1. Predict the outcome of the presidential election and the electoral college vote using the Linzer model to predict swing state outcomes in combination with presumed election results of "Red Wall" and "Blue Wall" states.
- 2. Predict whether the US Senate remains in Republican control by using an adaptation of the Linzer model and FiveThirtyEight Senate poll data for each state
- 3. Predict the outcomes of all 13 NC Congressional elections using Linzer model with input from FiveThirtyEight Senate poll data for North Carolina along with our model that predicts who will vote in North Carolina
- 4. The outcome of the NC Senate election and the associated uncertainty using the Linzer model.

Taking into account the anomaly that was the 2016 election, we chose to use the Linzer model to best account for slight and unexpected changes leading up to election day. In Section 2 of this report we will discuss datasets used for all tasks, including a brief exploratory data analysis. In Section 3, we formulate models and methodologies that answer all the research questions. Section 4 will present model diagnostic and sanity check validation of prediction. Then, in section 5 we present the major results of our analysis. Section 6 will be focused on conducting sensitivity analysis to test data imputation hypothesis and prior choices.

2. Data Source and EDA

2.1 Description of Data

In order to answer these questions, we used a total of four datasets: senate polls, house polls, 2020 US presidential election polls, and North Carolina voter registration history snapshot dataset. Both the senate and house polls dataset comes from the fivethirtyeight website [5], while the presidential election polls dataset comes from the Economist website [6].Senate and house polls have 38 variables and 4061 and 2655 observations respectively. The presidential election polls have 1447 observations and 17 variables which are: poll state, pollster, sponsor, start and end date, entry time, number of observations, population, method, Biden, Trump, Biden margin, other, undecided, URL, include, and notes.

The North Carolina voter history dataset contains information on the 2016 elections and about how voters voted (method, election, county, party affiliation, precinct). This will help us model and identify potential voters in 2020 (Interim report). On top of the created model, we used the 2020 voter registration snapshot dataset to identify potential voters, and use this information as additional input to model house polls results.

2.2 Exploratory Data Analysis

To get a basic understanding of the four datasets we were working with, we went through the data to understand what we were working with. In the senate and the house polls, we explored the variables of: methodology of the poll, the state in which the poll was conducted, which party the candidate was, and the percentage of the poll. In the presidential election polls, we explored similar variables of: method of poll and state, in addition to the margin between Biden and Trump. Because all three of these datasets had similar variables, we are able to compare them to each other. In Figure 1, we can see how the methods of polling were distributed between the house, senate, and presidential polls.

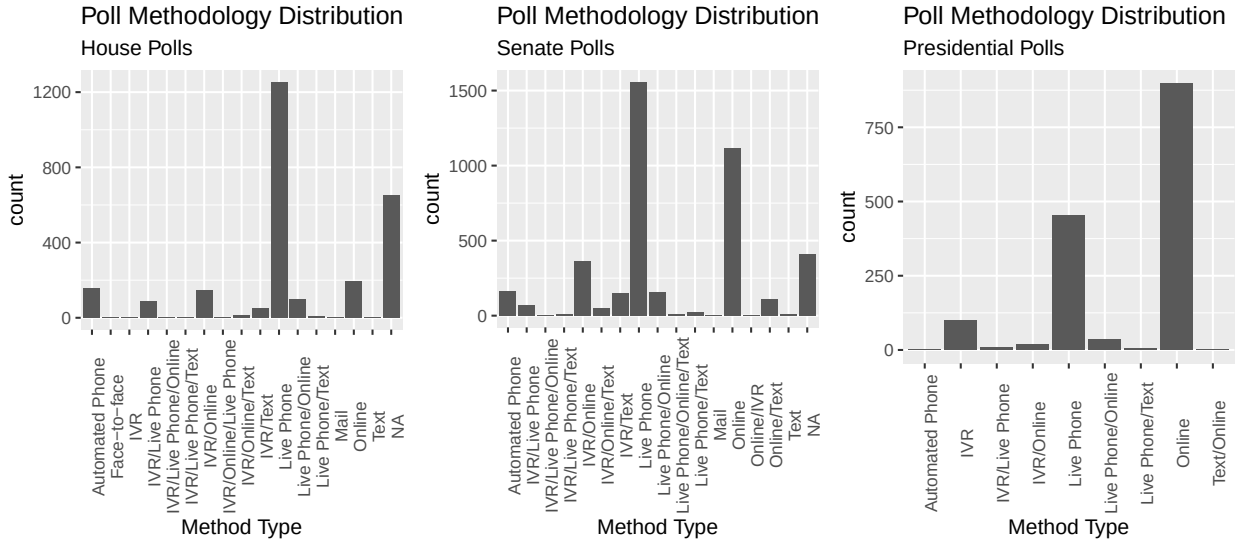


Figure 1: Poll Methodology

In Figure 2, we can see how each poll's distribution of which state was polled, and in Figure 3, we can see how the house and senate polls' candidate party are different from each other. Finally, in Figure 4, we can see the distribution of the margin between Biden and Trump.

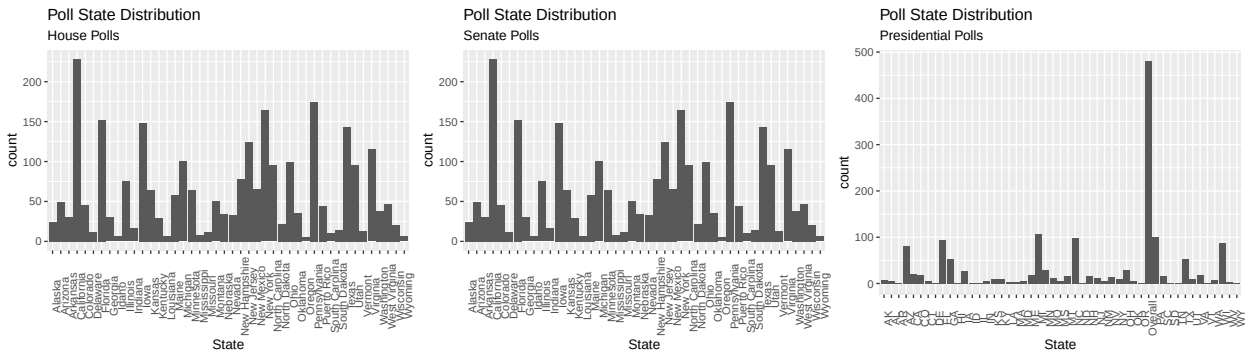


Figure 2: Polls State Distribution

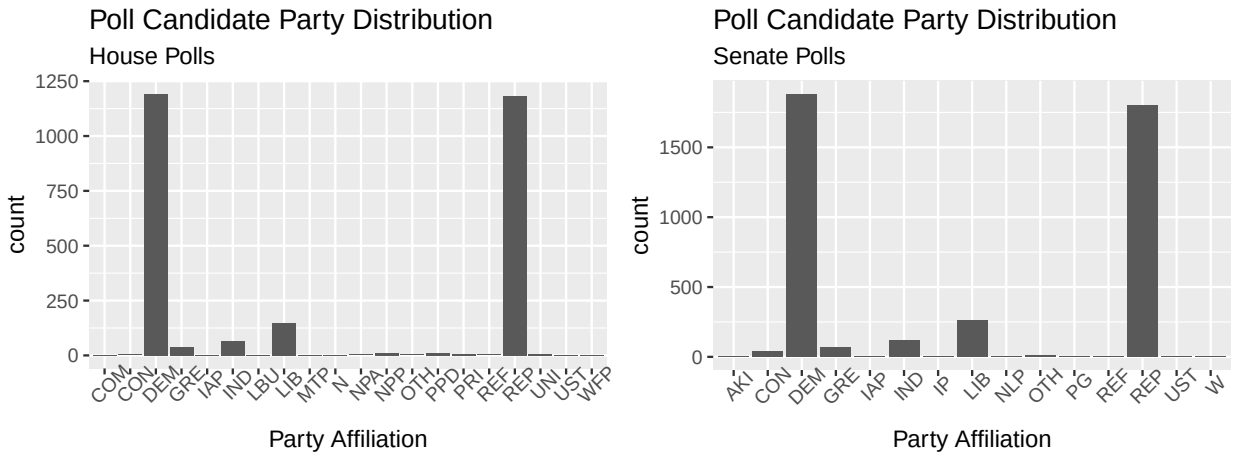


Figure 3: Candidate Party Polls

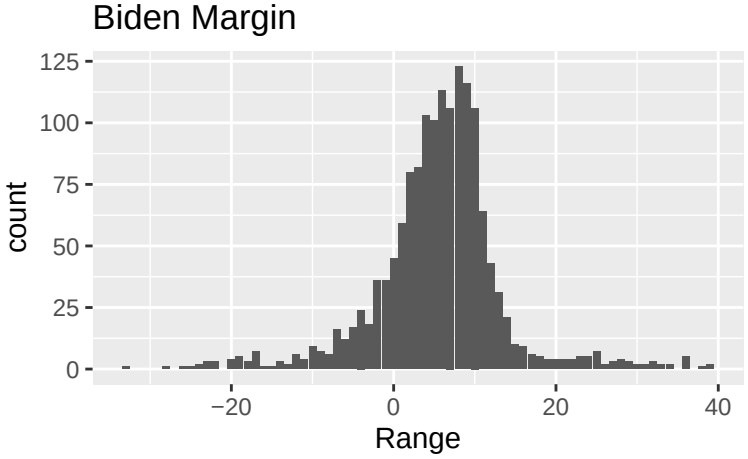


Figure 4: Biden Margin

To explore the North Carolina voter history dataset, we looked at variables such as method, county, and party affiliation of the voter. We can see in Figure 5 that the majority of North Carolina voters are from

two counties, Wake and Mecklenberg. Party affiliation and the method that the voters used are included in our appendix (——).

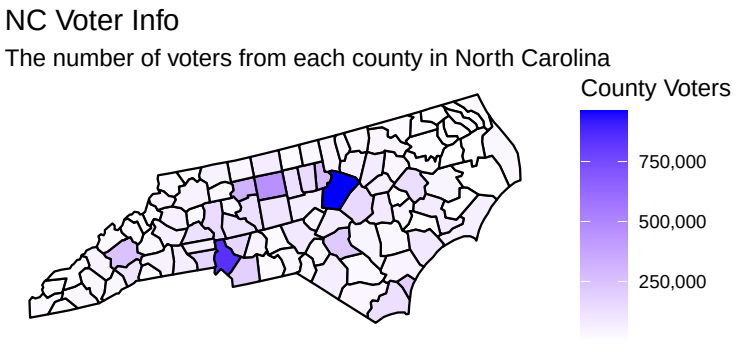


Figure 5: North Carolina Map of Voter Distribution by County

3. Model Formulation

3.1 Model for Question 1

In this section, we build a model to answer our first research question: predict the outcome of the presidential election and the electoral college votes. To do this, we use the Linzer model to predict swing state outcomes as well. To determine swing states, we've relied on worldpopulationreview's partisan index [7]. These states are Florida, Georgia, Iowa, North Carolina, Ohio, Texas, Arizona, Michigan, Minnesota, Nevada, New Hampshire, Pennsylvania, and Wisconsin. We tend to incorporate more states as swing states to avoid strong assumptions that any state is going to win. We follow the route of Linzer and build our model. Specifically, we want to model Joe Biden's percentage of EC votes on the state level. We believe that throughout time, each state's preference for Joe Biden randomly varies and follows a random walk process. Besides, each state's uncertainty of their attitude towards Joe Biden is also different statewide. We make the assumption that such uncertainty is constant over time. Therefore, we create the following extension of Linzer Model.

$$
\begin{aligned}
y_k &\sim N\left(\beta_{i[k]t[k]}, (\sigma_y^2)_{i[k]}\right) \\
\text{for } t > 1 : \beta_{it} &\sim N\left(\beta_{i,t-1}, (\sigma_\beta^2)\right) \\
\text{for } t = 1 : \beta_{i1} &\sim N\left(\mu_0, \sigma_0^2\right) \\
(\sigma_y^2)_{i[k]} &\sim \text{InvGamma}(\nu_y, \tau_y) \\
(\sigma_\beta^2) &\sim \text{InvGamma}(\nu_\beta, \tau_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\mu_0 &\sim N(50, 17) \\
\sigma_0^2 &\sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2}) \\
\nu_y &\sim \text{Uni}(0, 100) \\
\tau_y &\sim \text{Uni}(0, 100) \\
\nu_\beta &\sim \text{Uni}(0, 100) \\
\tau_\beta &\sim \text{Uni}(0, 100)
\end{aligned}
$$

Where k indexes the poll, i indexes the state, and t indexes the date. $i[k]$ represents the k^{th} poll's corresponding state index, and $t[k]$ represents the k^{th} poll's corresponding date index. We model samples from the economist poll data of 2020 and therefore are able to sample the posterior distribution of 1) Joe Biden's chance of winning, and 2) total electoral vote. The sampled posterior distribution y, which is the percentage of support within each state, will be calculated by comparing it to 50% to simulate the "winner takes all" procedure. After such adjustment, we can simply multiply the number of electoral college votes of each state to obtain a posterior distribution of electoral college votes for the entire nation. Thus, the total electoral vote can be obtained. By comparing this total to 270, we obtain the probability of Joe Biden being elected.

```
## [1] 303.4
```

3.2 Model for Question 2, 4

In this section, we're answering question 2 and 4 jointly: Predict whether the US Senate remains in republican control and predict the outcome of the NC Senate election. By using the above model with the same notation, and by switching Economist dataset to FiveThirtyEight senate polls data, the same inference procedure can be produced. This time, considering there are some states not running senator elections (Wisconsin, Ohio, Washington, Maryland, Pennsylvania, California, New York, Hawaii, Connecticut, Nevada, Indiana, North Dakota, Missouri, Utah, Vermont, Florida), we only model the states that are running senator elections. After modeling and predicting the Republicans' support percentage in each of these "electing states," we can predict whether the Republican or Democratic party wins the state majority in that state. This procedure also includes North Carolina. Furthermore, by summing the posterior samples of Republican senators in each state on election day, we can obtain the posterior distribution of whether the US Senate remains in Republican Party's control.

3.3 Model for Question 3

In this section, we answer research question 3: Predict the outcomes of all 13 NC Congressional elections. The dataset we used is the FiveThirtyEight house poll data. However, the dataset's sample sizes on each NC congressional district are small. This is due to the fact house polls happen less frequently compared to either senate polls or presidential polls. Besides, not all congressional districts are competitive because there is no term limit for representatives. FiveThirtyEight has only recorded poll history data for competitive districts including districts 2 (already conclusive), 3, 7, 8, 9, 11, 13. Thus, the correlation between non-competitive districts cannot be captured without inputting extra datasets. This section's modeling procedure will rely on the output of the interim reports' "who vote" results.

Using the model to predict who will vote in 2020, we ran this binary output model on all the registered voters in NC to predict whether they'll vote or not. In the NC voter registration profile 2020 snapshot dataset, there is a column indicating party affiliation of each potential voter [18]. Thus, we can predict the percentage of Republican voters for each congressional district and use this percentage as an imputed value for polls percentage. In this way, we're essentially imputing polling results y for non-competitive districts, which can be also fed into the Linzer Model defined below. A detailed description of how the imputation process work will be discussed after we introduce the Linzer model:

$$
\begin{aligned}
y_k \in Y &:= \{y | y \text{ observed in house poll or } y \text{ imputed}\} \\
y_k &\sim N\left(\beta_{i[k]t[k]}, (\sigma_y^2)_{i[k]}\right) \\
\text{for } t > 1 : \beta_{it} &\sim N\left(\beta_{i,t-1}, (\sigma_\beta^2)\right) \\
\text{for } t = 1 : \beta_{i1} &\sim N\left(\mu_0, \sigma_0^2\right) \\
(\sigma_y^2)_{i[k]} &\sim \text{InvGamma}(\nu_y, \tau_y) \\
(\sigma_\beta^2) &\sim \text{InvGamma}(\nu_\beta, \tau_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\mu_0 &\sim N(50, 17) \\
\sigma_0^2 &\sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2}) \\
\nu_y &\sim \text{Uni}(0, 100) \\
\tau_y &\sim \text{Uni}(0, 100) \\
\nu_\beta &\sim \text{Uni}(0, 100) \\
\tau_\beta &\sim \text{Uni}(0, 100)
\end{aligned}
$$

Where k indexes the polls or the imputed polls, i indexes the congressional district, and t indexes the polls' date. $i[k]$ represents the k^{th} poll's corresponding congressional district index, and $t[k]$ represents the k^{th} poll's corresponding date index.

Now we can start discussion of the imputation. To impute y, we are essentially "creating" more polls for those unobserved congressional districts. Suppose we want to create an y_{k+1} that happened on October 10 for district 1, a non-competitive district. We look into all the voters who've registered to vote in congressional district 1 prior to October based on the 2020 voter registration file snapshot dataset. Suppose n_{k+1} voters have registered. We use the model in our interim report to predict who among these n_{k+1} voters are likely to vote. Suppose there are n_{k+1}^* likely voters. Then, we calculate the percentage of voters who are Republicans. Such percentage is the value we've imputed for y_{k+1}, which will be served as an additional "poll" to be fed into the Linzer model. By imputing $y_{k+1}, y_{k+2} \cdots$ for all the non-competitive districts on each day following the above procedure, we've obtained a complete "polls record" for non-competitive districts. This can efficiently increase effective sample size for MCMC and also share mutual information on unobserved competitive districts via estimated correlation matrix.

Though efficient and reasonable, the imputation process above still makes strong assumptions. Thus, sensitivity analysis of imputed percentage will be further explored later in section 6.2.

4 Diagnostic and Validation

4.1 Model Diagnostic

Due to the fact that the parameter set is sparse – that is, β matrix has very few observations due to the lack of poll data, it is necessary to test the convergence of MCMC. We've run 6 MCMC chains on the input data for all the 3 models. Below in appendix (———) is some randomly selected model parameters' traceplot.

The reason for only illustrating some of them is that the models have too many parameters to keep track of. Thus, it is not feasible to illustrate them all. Trace plots show no significant convergent issues. Besides, the Gelman Rubin test has returned estimated Rhat to be 1.042, 1051, and 1.003, showing sufficient mixing. This holds for all the 3 MCMC models. Thus, we can deduce that MCMC has sufficiently converged.

4.2 Results Validation and Sanity Check

4.2.1 Presidential winner and EC vote share

Our model predicts that there is a 67.4% chance Biden will win the election. The median number of electoral college votes that our model simulates that Biden will win is 303 and the 95% confidence interval is (230, 387). We compared this to estimates from other publicly available models. FiveThirtyEight predicts Biden will win with an 89% probability and that Biden will win 348 electoral college votes [5]. The Economist model predicts that there is a 96% probability that Biden will win and the median number of EC votes that their model simulates Biden will win is 350 with a 95% confidence interval of (260, 420) [3]. Our model predicts the same outcome, but is slightly more conservative which we think is appropriate given how "sure" the 2016 models were that Clinton would win and how wrong they turned out to be in reality. In terms of swing states that will likely determine the EC vote share, our model predictions differ from those of FiveThirtyEight and the Economist. Our model and both of theirs predict that AZ, NC, MI, WI, MN, FL, PA, NV and NH are the swing states that Biden will win. This makes sense because these swing states are the ones that are the least contentious and have historically (before 2016), voted democrat more often [8]. Our models along with theirs predict that IA, TX, and OH are swing states that Trump will win, which makes sense because in 2016 these were the swing states that Trump won by the largest margin [9]. Our models differ since we predict that Trump will also win the EC votes from GA. This is appropriate because Trump won this state by a larger margin in 2016 than the other swing states that we predict Biden will win (NYT). Our model in general makes these predictions with less certainty (probabilities closer to 50%) than the FiveThirtyEight and Economist models. We believe this is appropriate that our model is more conservative on Biden winning EC votes from swing states because only 30.3 % of Republicans have done early voting as of 10/31/2020 [5]. In these states specifically, Republicans have shown up to the polls closer towards the end of early voting and the gap has narrowed [10].

4.2.2 Partisan Control of the Senate

There are 35 seats up for reelection and the Democrats need to win 3-4 more seats to take control. 16 states have no elections this cycle. Two states have special elections, which means that they were not supposed to have one of their seats up for election this year, but due to special circumstances, they now do. Arizona is having a special election because the incumbent, John McCain (R) died during his term [15]. Georgia is also having a special election because the incumbent, Johnny Isakson (R), resigned after he was diagnosed with Parkinson's [16]. Our model predicts that Republicans have a higher likelihood of winning Senate elections in AZ, NC, IA, GA, TX, AK, AL, KY, WY, MT, KS, ME, SC, MS, TN, NE, AR, OK, OR, and ID. We differ from the FiveThirtyEight predictions because they claim WV and LA will go red and AZ, NC, ME, and OR will go blue. In LA, there has been recent talk of a potential run-off election given how close the polls have shown the candidates to be so our prediction makes sense [19]. We believe that our predictions that ME, OR, AZ, and NC will go red are reasonable because many undecided voters are leaning Republican and margins between candidates have narrowed this week [17] [10]. Overall, we predict that there is a 20.03% probability that the senate will remain in Republican control which is reasonable given that the democrats only need to flip a few more seats to achieve this and it is close to FiveThirtyEight's 25% probability [5].

4.2.3 NC Senate Election

Our model predicts that the median amount of votes that Thom Tillis (R) will win is 48.27% with a 95% confidence interval of (45.65%, 50.71%). Our model estimates that his probability of winning is 57.8%. Note

that Tillis does not need to win over 50% of votes to win the election because of the votes that go towards third parties. He just needs to win more votes than Cunningham. Our estimates make sense because Tillis is the incumbent, but the race is very close due to controversy surrounding his outspoken support for Trump and his opposition to the Affordable Care Act [11]. Tillis has joined Trump at many NC rallies and has support from Vice President Pence. This means that the race will likely follow the results of the General Election (i.e. Trump's fate in NC will likely determine Tillis's) [11]. FiveThirtyEight's latest prediction is that Cunningham will win and predicts that Tillis will gain 46% of the vote with a 4% chance of being elected and the Economist's latest prediction has Tillis winning 48.4% of the vote with a 27% probability of winning [5] [3]. However, we believe our model is reasonable due to the recent scandal leaking private text messages revealing that Cunningham was having an affair. The gap between them has begun to close [12].

4.2.4 Congressional District election for the 13 NC districts

All districts, except the 11th, have incumbents running for reelection. In district 2, the opponent has dropped out so the representative is automatically Republican George Holding [12]. The 11th district has been 9 points more Republican than the national average in the last 2 elections [12]. However, this district historically has followed the general election. Ballotpedia has combined various polls to predict that the district is leaning Republican and that Mark Meadows (R) will get 59.2% of the vote share [12]. The median prediction from our model simulations is that Meadows will gain 48.41% of votes with a 36.57% probability of winning. We think this is reasonable given how dependent the district is on the general election. We predict that districts 1, 4, 5, 6, 7, 10, and 12 are almost certain to elect their incumbents. This makes sense for these districts because in these districts the incumbent has been reelected several times [12]. That leaves districts 9, 8, 3, and 13. District 9 is a close race because the current incumbent, Dan Bishop (R) , sponsored the infamous House Bill 2, otherwise known as the "bathroom bill," which told transgender people to use the bathrooms that matched their birth sex. The median prediction from our model simulations is that Bishop will gain 49.55% of votes with a 40.17% probability of winning. This makes sense given how much bad press Bishop has gotten since he supported the "bathroom bill" [13]. District 8 votes 5% points more Republican than the national average, but the district has recently gone through redistricting which will make many voters unfamiliar to incumbent Richard Hudson (R) [12] [14]. Given this, the race will be close, so our simulations predict that Hudson will win 49.55% of the vote share and that he has a 64.33% probability of being elected. In district 3 and 13, there is a 66.77% and 68.97% probability that the Republican incumbent wins, respectively. Even though these Republican incumbents are favored to win, these districts typically follow the partisan results of the general election [12]. Overall, our prediction estimates make sense based on context.

5 Modeling Result

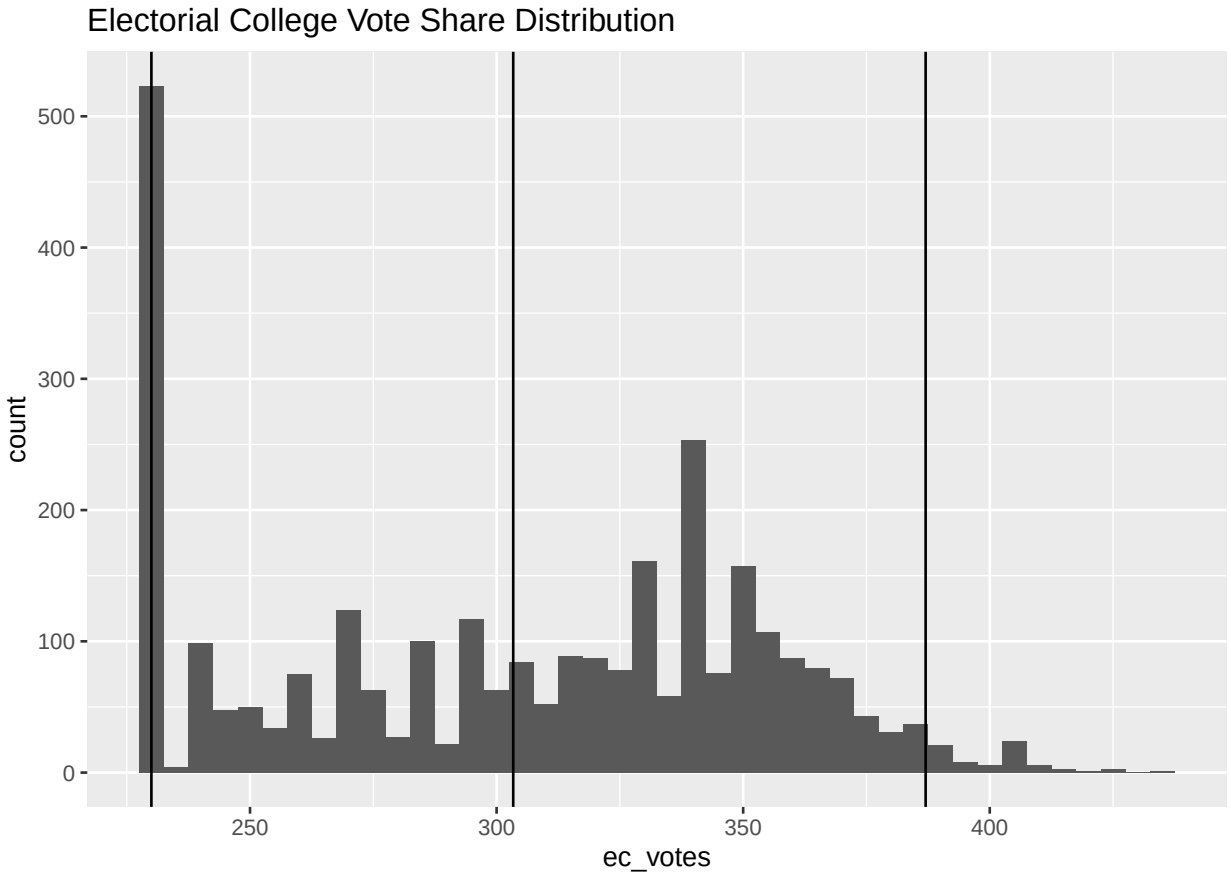
5.1 Presidential Election and Electoral Vote

Our model predicts that Biden will win with a percentage of 67.4% and the electoral college vote will be 303 (rounded from 303.4) with the 95% confidence interval being (230, 387) votes as shown in Table 1.

```
## [1] "biden wins 0.674"
```

Table 1: EC Vote Total (Biden)

	x
2.5%	230
97.5%	387



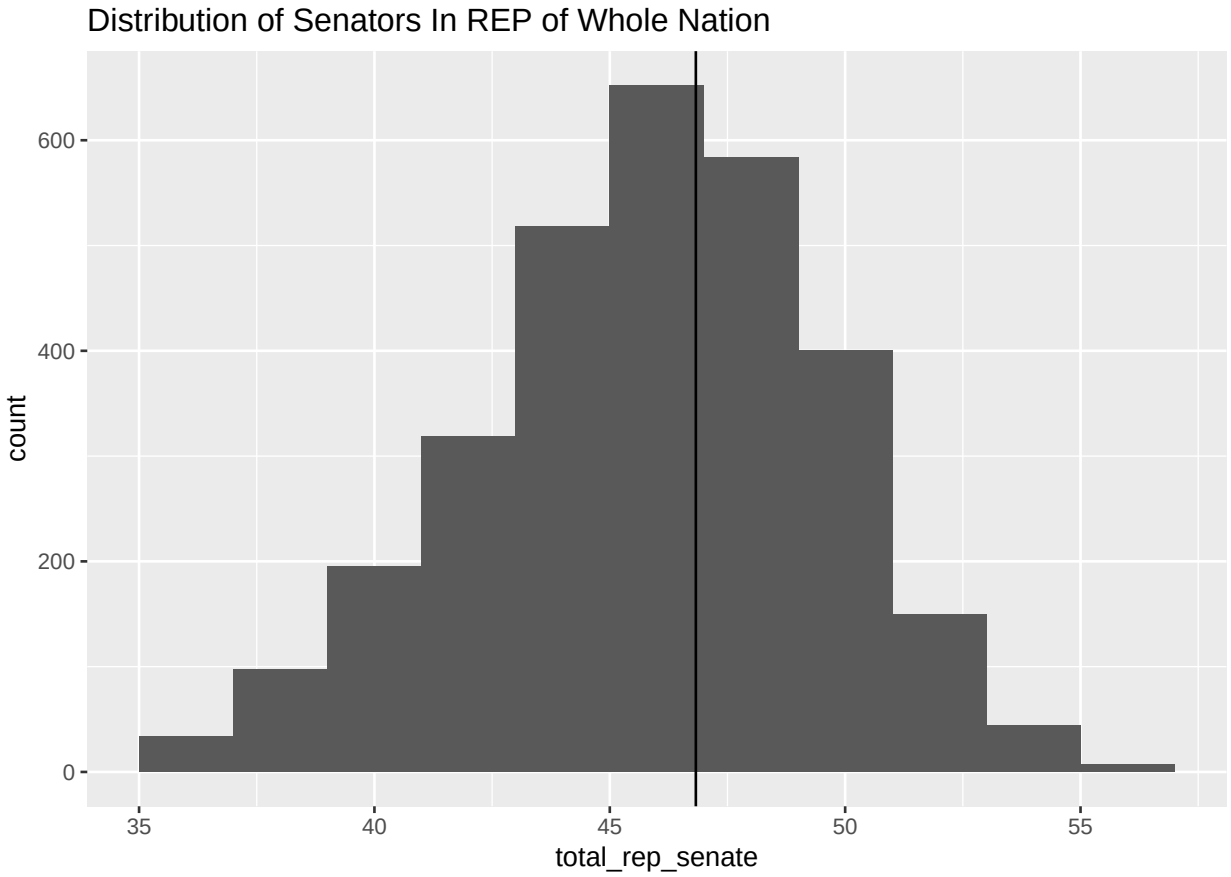
5.2 Whether Senate Remains in Republican Control

Our model predicts that the Republican Party will remain in control of the Senate with only a probability of 20%, meaning it is more likely that the Democratic Party will be in control of the Senate.

```
## [1] "republican control senate 0.200333333333333"
```

Table 2: House All Congressional District Winning Probability (REP)

	Official Model	Informative Prior	Imputation Perturbation
District 9	0.4017	0.1960	0.2727
District 11	0.3657	0.3963	0.4260
District 8	0.6433	0.2310	0.3357
District 3	0.6677	0.4683	0.6873
District 13	0.6897	0.4643	0.6010
District 2	0.7193	0.7687	0.4877
District 7	1.0000	0.6803	0.6437
district 1	0.0000	0.0000	0.0000
district 4	0.0000	0.0000	0.0000
district 5	1.0000	1.0000	1.0000
district 6	0.0000	0.0000	0.0000
district 10	1.0000	1.0000	1.0000
district 12	0.0000	0.0000	0.0000



5.3 13 NC Congressional District Result

In North Carolina, Tom Tillis is predicted to win the Senate seat with a probability of 57.8%.

Table 4: North Carolina Vote Share Interval Estimate

	x
2.5%	45.65
50%	48.27
97.5%	50.71

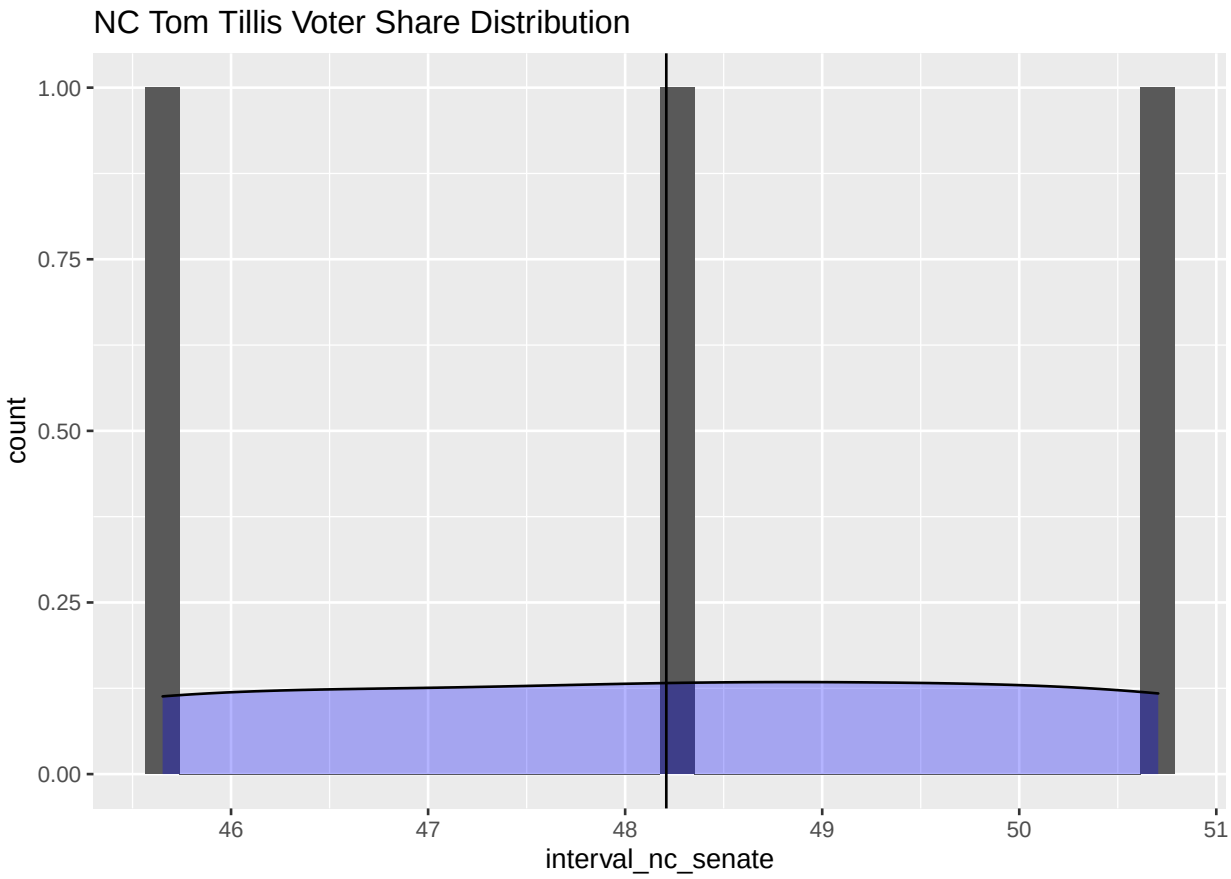
Table 3: House All Congressional District Winning Probability (REP)

	Official Model			Informative Prior			Imputation Perturbation		
	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
District 9	46.10	49.55	53.44	45.58	48.66	51.73	45.73	48.97	51.97
District 11	41.47	48.41	55.71	44.63	48.80	61.46	44.57	49.45	53.98
District 8	40.74	54.34	71.66	38.52	47.43	53.10	39.25	45.08	57.74
District 3	39.70	54.16	61.31	40.64	49.68	56.30	45.03	52.17	64.02
District 13	46.52	52.91	60.37	41.38	49.44	58.86	45.72	50.91	58.35
District 2	41.23	52.33	61.45	44.85	52.88	58.20	43.50	49.88	57.10
District 7	53.73	57.52	67.91	42.90	57.15	69.12	32.32	51.88	56.09
district 1	26.12	26.83	27.59	26.56	26.84	27.10	26.66	26.84	27.01
district 4	22.53	22.78	23.03	22.45	22.79	23.13	22.57	22.78	23.00
district 5	68.37	68.55	68.74	68.38	68.56	68.74	68.38	68.56	68.73
district 6	45.59	45.72	45.85	45.59	45.72	45.85	45.60	45.72	45.85
district 10	69.72	70.02	70.30	69.53	70.02	70.47	69.85	70.02	70.18
district 12	36.11	36.27	36.43	36.12	36.27	36.41	36.12	36.27	36.41

5.4 NC Senator Election

In District 1 of North Carolina, G. K. Butterfield (D) is predicted to win. In District 2, Alan Swain (R) is predicted to win, and in District 3, Gregory Murphy (R) will win. District 4 is predicted to elect David Price (D). District 5 is predicted to elect Virginia Foxx (R). In District 6, Kathy Manning (D) is predicted to be elected. In District 7, David Rouzer (R) is predicted to win, and in District 8, Richard Hudson (R) is predicted to win. In District 9, Cynthia Wallace (D) is predicted to win, and in District 10, Patrick T. McHenry (R) is predicted to win. Morris Davis (D) is predicted to win District 11. In District 12, Alma Adams (D) is predicted to win. Finally, in District 13, Ted Budd (R) is predicted to win. The probability and confidence intervals for the republican party winning the congressional district for each of North Carolina's 13 Congressional Districts are shown in Table 3 below.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## [1] "Tome Tillis Win Prob 0.578"
```

6 Sensitivity Analysis

6.1 Informative Prior Check

As indicated in the model above in section 3, compared to the size of the β matrix, the total number of samples to y_k is comparably small. Thus, we're concerned that variation of prior values of parameters can lead to drastically different results. As we've already defined in section 3, we've set the prior mean to be 50%. This prior mean value is arguably the most reasonable value. Thus, this sensitivity check section will not test sensitivity of μ_0. However, all the other parameters, including $\nu_y, \nu_\beta, \tau_y, \tau_\beta, \sigma^2$ control the prior population (inverse of uncertainty) of belief that the vote share percentage is 50%. Augmenting their values will make the prior less informative, whereas decreasing their values will make the Linzer model's posterior harder to deviate from generating 50% vote share percentage. Thus, we're testing the sensitivity to such uncertainty preset in our model by comparing the set of prior defined in section 3 to the following set of priors:

$$\mu_0 \sim N(50, 1)$$
$$\sigma_0^2 \sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2})$$
$$\nu_y \sim \text{Uni}(0, 1)$$
$$\tau_y \sim \text{Uni}(0, 1)$$
$$\nu_\beta \sim \text{Uni}(0, 1)$$
$$\tau_\beta \sim \text{Uni}(0, 1)$$

The result is also attached in Table (——) (——) (——) (——) (——) (——) in the appendix. In all these tables, the "Official Model" column gives point or interval estimates using the prior in section 3, the model section, whereas the "Informative Prior" column gives all the same results using the above set of prior. Based on the observation, though the range of uncertainty parameters have been adjusted to 2 digits of magnitude, both estimates are still very similar. Thus, this justifies 2 facts:

- The sample size in the Linzer Model is sufficient in providing enough information (rather than only drawing information from the priors) to inform posterior distributions
- Our model generates convergent results that are insusceptible from prior choice.

Thus, our model outputs are valid and trustworthy.

6.2 Imputation Perturbation Check

As mentioned before in section 3.3, the congressional district prediction utilizes manual imputation to generate forged poll data y for noncompetitive districts. We're highly concerned about whether such imputation is valid and whether any trivial deviation from truth will introduce overwhelming errors. Thus, we've also conducted the perturbation check on imputed values.

The perturbation is conducted as follows. For all the imputed y, we add independent yet identical Gaussian noise to them.

$$y_{\text{perturbed imputation}} := y_{\text{original imputation}} + \epsilon \quad \epsilon \sim N(0, \kappa^2)$$

We're concerned whether 10% deviation of imputed vote share percentage from the truth vote share percentage will generate completely different results. Therefore, we chose $\kappa = 5\%$, that 10% range is plus or minus 2 standard deviations from the mean. This is the perturbation process we've conducted.

Similarly, before and after perturbation, the results are all attached in Table (——) (——) (——) (——) (——) (——). This time, the "Imputation Perturbation" column contains all the estimated results. We still observe no major difference between the estimated probability or vote share interval estimates. Thus, small errors introduced in or imputation won't affect the final prediction too much. Thus, the results generated upon imputed values are credible.

Bibliography

[1] Boyarsky, B. (2019). Are Polls Reliable? Retrieved November 02, 2020, from https://blueprint.ucla.edu/feature/are-polls-reliable/

[2] Shapiro, W. (2019, June 21). The Polling Industry Is in Crisis. Retrieved November 02, 2020, from https://newrepublic.com/article/154124/polling-industry-crisis

[3] How The Economist presidential forecast works. (n.d.). Retrieved November 02, 2020, from https://projects.economist.com/us-2020-forecast/president/how-this-works?fbclid=IwAR2DIgVtL4uVjyKiVrpj9f3KMiUyOzvmsTK

[4] Linzer, D. A. (2013). Dynamic Bayesian Forecasting of Presidential Elections in the States. Journal of the American Statistical Association, 108(501), 124-134. doi:10.1080/01621459.2012.737735

[5] DataDhrumil. (2020, November 02). U.S. Senate Polls. Retrieved November 02, 2020, from https://projects.fivethirtyeight.com/polls/senate/

[6] How The Economist presidential forecast works. (2020). Retrieved November 02, 2020, from https://projects.economist.com/us-2020-forecast/president/how-this-works

[7] Blue States 2020. (n.d.). Retrieved November 02, 2020, from https://worldpopulationreview.com/state-rankings/blue-states

[8] Staff, N. (2016, November 2). A recent voting history of the 15 Battleground states. Retrieved November 02, 2020, from https://constitutioncenter.org/blog/voting-history-of-the-15-battleground-states

[9] 2016 Presidential Election Results: Donald J. Trump Wins. (2017, August 9). Retrieved November 02, 2020, from https://www.nytimes.com/elections/2016/results/president

[10] Riccardi, N., & Kastanis, A. (2020, October 25). Early vote total exceeds 2016; GOP chips at Dems' advantage. Retrieved November 02, 2020, from https://apnews.com/article/election-2020-donald-trump-politics-florida-elections-509ad83f6d40e08fb715da44548f62e0

[11] Arkin, J. (2020, November 01). Tillis scrambles, Cunningham lies low in closing days of N.C. Senate race. Retrieved November 02, 2020, from https://www.politico.com/news/2020/11/01/tillis-cunningham-north-carolina-senate-race-final-days-433783

[12] United States Senate election in North Carolina, 2020. (n.d.). Retrieved November 02, 2020, from https://ballotpedia.org/United_States_Senate_election_in_North_Carolina,_2020

[13] Dalesio, E. P. (2019, May 11). North Carolina 'bathroom bill' sponsor bidding for US House. Retrieved November 02, 2020, from https://apnews.com/article/44b86b14141a41a6a3efc857273eb7d3

[14] Anderson, B., & Robertson, G. D. (2020, October 30). Republicans on defense in North Carolina congressional races. Retrieved November 02, 2020, from https://apnews.com/article/election-2020-donald-trump-legislature-hudson-north-carolina-798c7de60ca763fe9686c69d39b85590

[15] Arizona race could give Democrats extra Senate seat for supreme court fight. (2020, September 20). Retrieved November 02, 2020, from https://www.theguardian.com/us-news/2020/sep/20/arizona-democrats-senate-race-supreme-court-ruth-bader-ginsburg

[16] Everett, B. (2019, August 28). Sen. Johnny Isakson to resign at end of the year. Retrieved November 02, 2020, from https://www.politico.com/story/2019/08/28/sen-johnny-isakson-to-resign-at-end-of-the-year-1476655

[17] Shepherd, M., Andrews, C., & Piper, J. (2020, November 02). New polls show Maine's US Senate race is still wracked with uncertainty. Retrieved November 02, 2020, from https://bangordailynews.com/2020/11/02/politics/daily-brief/new-polls-show-maines-us-senate-race-is-still-wracked-with-uncertainty/

[18] Dl.ncsbe.gov. 2020. [online] Available at: https://dl.ncsbe.gov/?prefix=data/Snapshots/ [Accessed 1 November 2020].

[19] https://www.theadvocate.com/baton_rouge/news/politics/elections/article_4923a528-1c94-11eb-aa24-5bb599e01390.html

Appendix A

Model 1, to Model Presidential Election Outcome and EC Vote Share

$$
\begin{aligned}
y_k &\sim N\left(\beta_{i[k]t[k]}, (\sigma_y^2)_{i[k]}\right) \\
\text{for } t > 1 : \beta_{it} &\sim N\left(\beta_{i,t-1}, (\sigma_\beta^2)\right) \\
\text{for } t = 1 : \beta_{i1} &\sim N\left(\mu_0, \sigma_0^2\right) \\
(\sigma_y^2)_{i[k]} &\sim \text{InvGamma}(\nu_y, \tau_y) \\
(\sigma_\beta^2) &\sim \text{InvGamma}(\nu_\beta, \tau_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\mu_0 &\sim N(50, 17) \\
\sigma_0^2 &\sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2}) \\
\nu_y &\sim \text{Uni}(0, 100) \\
\tau_y &\sim \text{Uni}(0, 100) \\
\nu_\beta &\sim \text{Uni}(0, 100) \\
\tau_\beta &\sim \text{Uni}(0, 100)
\end{aligned}
$$

Where k index the polls, i index the states, and t index the date. $i[k]$ represents the k^{th} poll's corresponding state index, and $t[k]$ represents the k^{th} poll's corresponding date index. We input this model into MCMC sampler and sample $\beta_{i,T} \forall i$ where T denotes the last day of election. Thus, by taking $\beta_{i,T} \forall i$ of all MCMC iterations, we can obtain the empirical posterior distribution of electoral college vote share of all modeled swing states (Florida, Georgia, Iowa, North Carolina, Ohio, Texas, Arizona, Michigan, Minnesota, Nevada, New Hampshire, Pennsylvania, and Wisconsin).

For each MCMC iteration, all the $\beta_{i,T} \forall i$ are compared against 50%. If it is bigger than 50%, it means that in the corresponding state, Biden has won all the electoral votes of its state, therefore taking 100% percent of the electoral votes of the state. Otherwise, Biden takes 0%. Therefore, we create a list of 0% and 100%, used to dot product with the number of electoral votes for each state, then plus the total electoral votes of all the "Blue Wall" states, we end up with the total number of Electoral Vote for Biden in this MCMC iteration. Doing the above calculation for all MCMC iteration, we can obtain the posterior distribution of EC votes that Biden wins. From which, we can calculate the probability that Biden wins the election, which equals the number of EC votes bigger than 270 over the total MCMC iterations.

Model 2, to Model Senator and NC Senator Election

$$
\begin{aligned}
y_k &\sim N\left(\beta_{i[k]t[k]}, (\sigma_y^2)_{i[k]}\right) \\
\text{for } t > 1 : \beta_{it} &\sim N\left(\beta_{i,t-1}, (\sigma_\beta^2)\right) \\
\text{for } t = 1 : \beta_{i1} &\sim N\left(\mu_0, \sigma_0^2\right) \\
(\sigma_y^2)_{i[k]} &\sim \text{InvGamma}(\nu_y, \tau_y) \\
(\sigma_\beta^2) &\sim \text{InvGamma}(\nu_\beta, \tau_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\mu_0 &\sim N(50, 17) \\
\sigma_0^2 &\sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2}) \\
\nu_y &\sim \text{Uni}(0, 100) \\
\tau_y &\sim \text{Uni}(0, 100) \\
\nu_\beta &\sim \text{Uni}(0, 100) \\
\tau_\beta &\sim \text{Uni}(0, 100)
\end{aligned}
$$

Where k index the polls, i index the states, and t index the date. $i[k]$ represents the k^{th} poll's corresponding state index, and $t[k]$ represents the k^{th} poll's corresponding date index. We input this model into MCMC sampler and sample $\beta_{i,T} \forall i$ where T denotes the last day of election. Thus, by taking $\beta_{i,T} \forall i$ of all MCMC iterations, we can obtain the empirical posterior distribution of senator vote share of all modeled states (all states excluding Wisconsin, Ohio, Washington, Maryland, Pennsylvania, California, New York, Hawaii, Connecticut, Nevada, Indiana, North Dakota, Missouri, Utah, Vermont and Florida).

For each MCMC iteration, all the $\beta_{i,T} \forall i$ are compared against 50%. If it is bigger than 50%, it means that in the corresponding state, Republican has won all the senator spots of its state, therefore taking 100% percent of the senator spots of the state. Otherwise, Republican take 0%. Therefore, we create a list of 0% and 100%, used to dot product with the number of available senator spots for each state, then plus the total existing republican senators of all the non-modeled states, we end up with the total number of senators for Republican Party in this MCMC iteration. Doing the above calculation for all MCMC iteration, we can obtain the posterior distribution of the number of senators that Republican have in the Senate after this election. From which, we can calculate the probability that Republican controls the Senate, which equals the number of Republican Senators bigger than 50 over the total MCMC iterations.

Model 3, to model 13 Congressional District

Similarly, we have the same model.

$$
\begin{aligned}
y_k \in Y &:= \{y | y \text{ observed in house poll or } y \text{ imputed}\} \\
y_k &\sim N\left(\beta_{i[k]t[k]}, (\sigma_y^2)_{i[k]}\right) \\
\text{for } t > 1 : \beta_{it} &\sim N\left(\beta_{i,t-1}, (\sigma_\beta^2)\right) \\
\text{for } t = 1 : \beta_{i1} &\sim N\left(\mu_0, \sigma_0^2\right) \\
(\sigma_y^2)_{i[k]} &\sim \text{InvGamma}(\nu_y, \tau_y) \\
(\sigma_\beta^2) &\sim \text{InvGamma}(\nu_\beta, \tau_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\mu_0 &\sim N(50, 17) \\
\sigma_0^2 &\sim \text{InvGamma}(\frac{1}{2}, \frac{1}{2}) \\
\nu_y &\sim \text{Uni}(0, 100) \\
\tau_y &\sim \text{Uni}(0, 100) \\
\nu_\beta &\sim \text{Uni}(0, 100) \\
\tau_\beta &\sim \text{Uni}(0, 100)
\end{aligned}
$$

Where k indexes the polls or the imputed polls, i indexes the congressional district, and t indexes the polls' date. $i[k]$ represents the k^{th} poll's corresponding congressional district index, and $t[k]$ represents the k^{th} poll's corresponding date index.

However, we're creating more synthetic y into the house_polls dataset to enlarge. These imputed $Y := \{y | y \text{ observed in house poll or } y \text{ imputed}\}$. We only impute y values for noncompetitive districts including district 1, 4, 5, 6, 10, 12. The imputation procedure follows as below:

First, we use the trained model from the interim report to predict who are likely to vote in the 2020 voter registration snapshot dataset. Below is the model:

$$
\begin{aligned}
y_i &= \text{Bernoulli}(p_i) \\
logit(p_i) &= \beta_0 + \beta_{\text{ status code}} \text{ status code}_i + \beta_{\text{ race}} \text{ race}_i + \beta_{\text{ age}} \text{ age}_i + \beta_{\text{ drivers licence}} \text{ drivers license}_i \\
&\quad + \beta_{\text{ congressional district}} \text{ congressional district}_i + \beta_{\text{ ethnic}} \text{ ethnic}_i \\
&\quad + \beta_{\text{race : age}} \text{ race}_i \times \text{age}_i + \beta_{\text{congressional district : age}} \text{ congressional district}_i \times \text{age}_i
\end{aligned}
$$

For each registered voter, predict whether they vote or not using the model above. After that, we only take the likely voters for our further imputation analysis.

Then, for all $t < T$, where T is the total number of days prior to election days we're using for modeling (we set T=100), do the following. For each noncompetitive district, indexed as i, Count the total number of voters who registered before day t, denoted as $V_{t,1}$. Also, count how many are affiliated as republican voters among these $V_{t,1}$ people. Denote this number as $r_{t,1}$. Therefore, we impute $y = r_{t,1}/V_{t,1}$, which is a "synthetic" poll data for a non-competitive district on a specific date. Finally, just add this y into Y.

We only impute the above value for day t prior to 3 days before final election to enable some uncertainty.

Appendix B

List of data sets/data sources

- 1. House Polls

This dataset comes from the fivethirtyeight website and contains candidate party information, method of poll, and more information on the poll for the House of Representative polls.

- 2. Senate Polls

This dataset comes from the fivethirtyeight website and contains candidate party information, method of poll, and more information on the poll for the Senate polls.

- 3. Presidential Election Polls 2020

This dataset comes from the fivethirtyeight website and contains information on biden margin, mode of poll, number of observations, pollster information, and state.

- 4. NC Voter History Statewide Small (VR_Snapshot)

NC Voter History Statewide Small provides voter history information for North Carolina voters including their voting method, which election, and registered party. NCSBE Snapshots

Model 1, to Model Presidential Election Outcome and EC Vote Share

Dataset 3 [Presidential Election Polls 2020]

- k `polls` variable
- i index `state`, with $i[k]$ the poll's state index
- t index `date`, with $t[k]$ the poll date index

Model 2, to Model Senator and NC Senator Election

Dataset 2 [Senate Polls] and Dataset #4 [NC Voter Statewide Small]

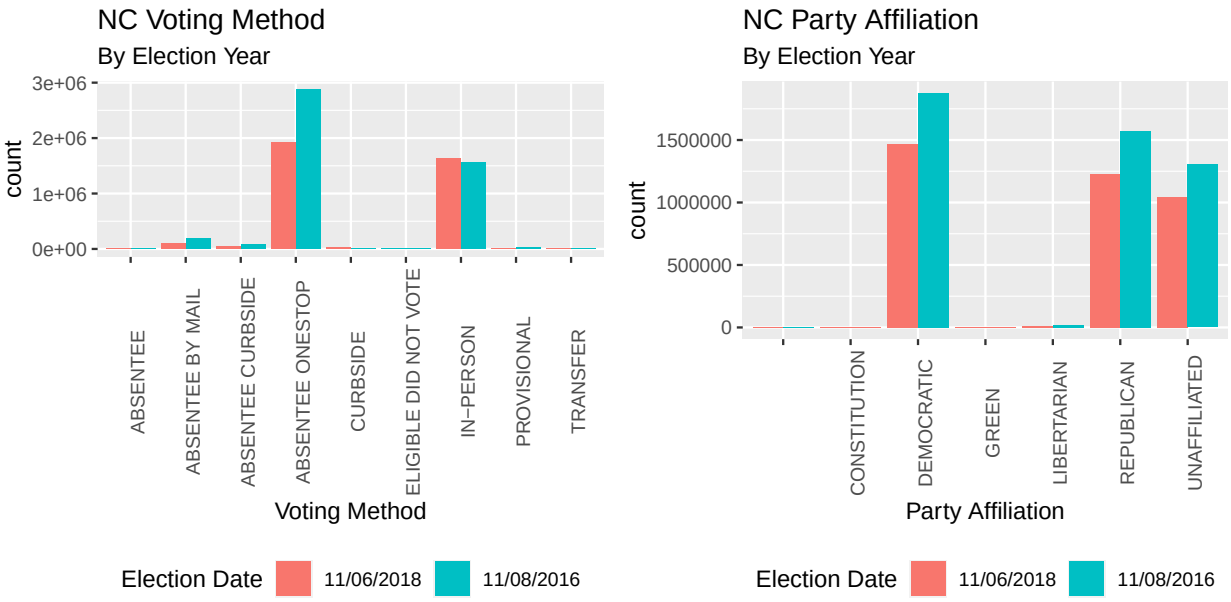
- k `polls` variable
- i index `state`, with $i[k]$ the poll's state index
- t index `date`, with $t[k]$ the poll date index

Model 3, to Model 13 Congressional District

Dataset 4 [NC Voter Statewide Small (NCSBE Snapshots)]

- k `polls` variable
- i index `congressional_district`, with $i[k]$ the poll's congressional district index
- t index `date`, with $t[k]$ the poll date index

Appendix




```
## Warning: Removed 2044 rows containing missing values (geom_point).
```

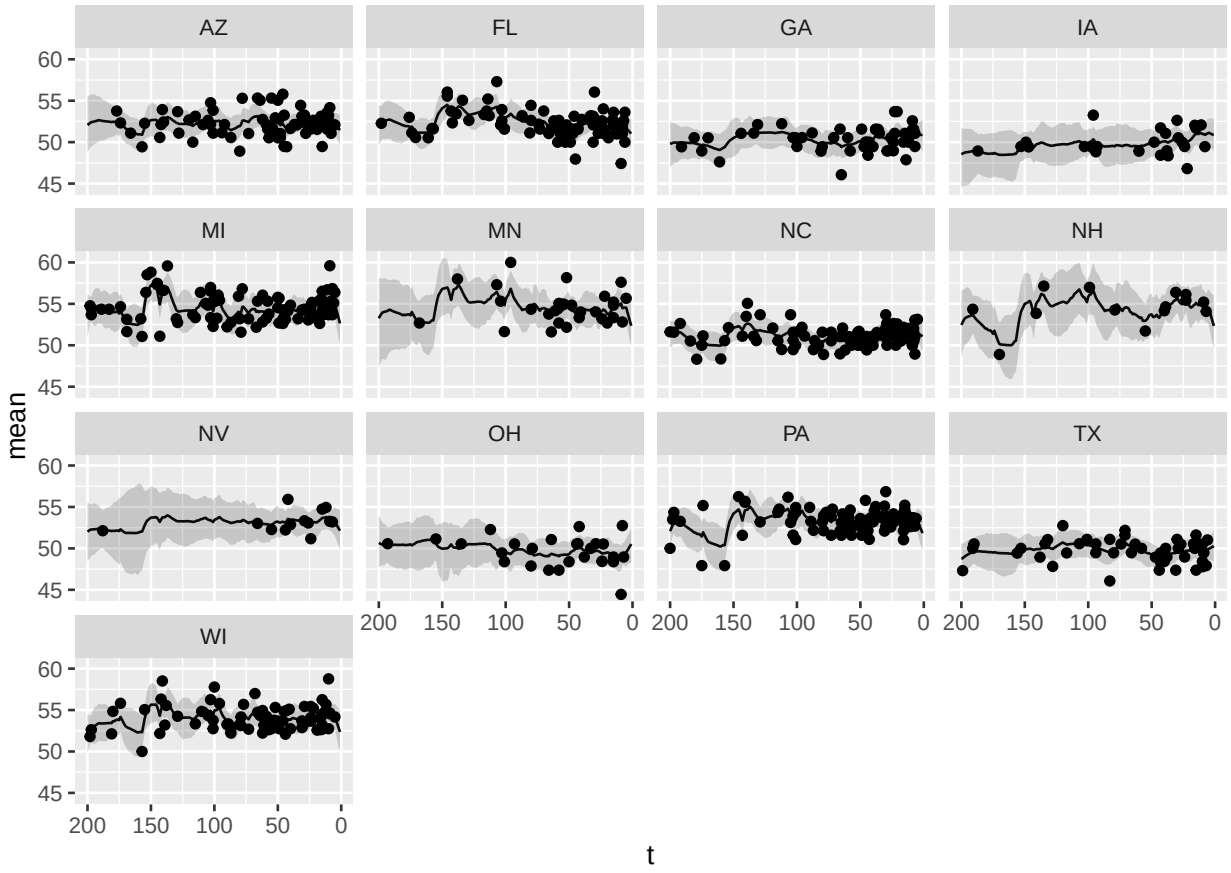

AZ
FL
GA
IA
MI
MN
NC
NH
NV
OH
PA
TX
WI
mean
60
55
50
45
200
150
100
50
0
t

Table 5: Swing State Winning Probability (Biden)

	Official Model	Informative Prior	Imputation Perturbation
AZ	0.6860	0.6677	0.7753
NC	0.5443	0.5367	0.5643
MI	0.8897	0.9047	0.9467
WI	0.8517	0.8730	0.9173
MN	0.8313	0.8457	0.8867
TX	0.2127	0.2497	0.1847
FL	0.5093	0.5007	0.5937
PA	0.7557	0.7370	0.8123
GA	0.4390	0.4093	0.3990
OH	0.3317	0.3143	0.2817
NV	0.8210	0.8550	0.8807
NH	0.8067	0.8210	0.8843
IA	0.4363	0.4813	0.4867

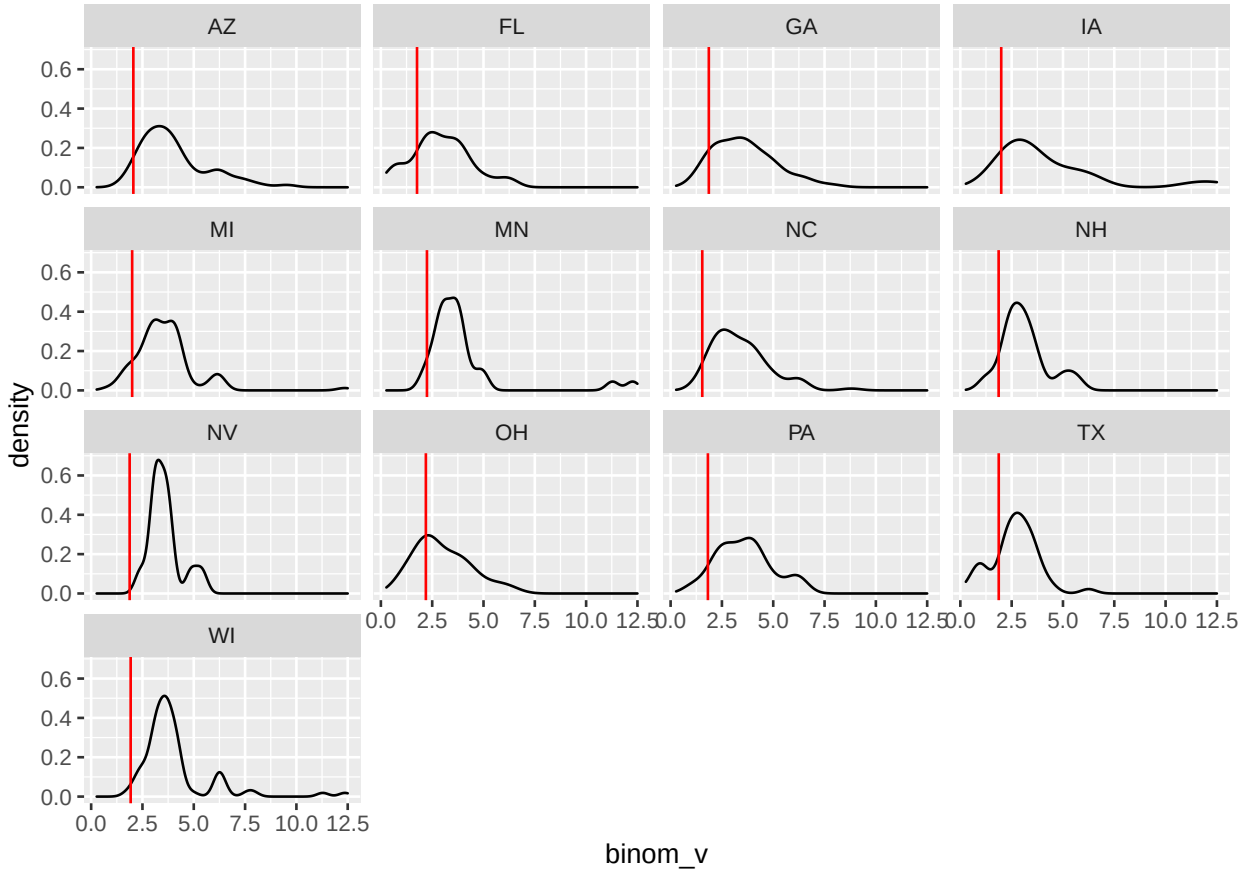
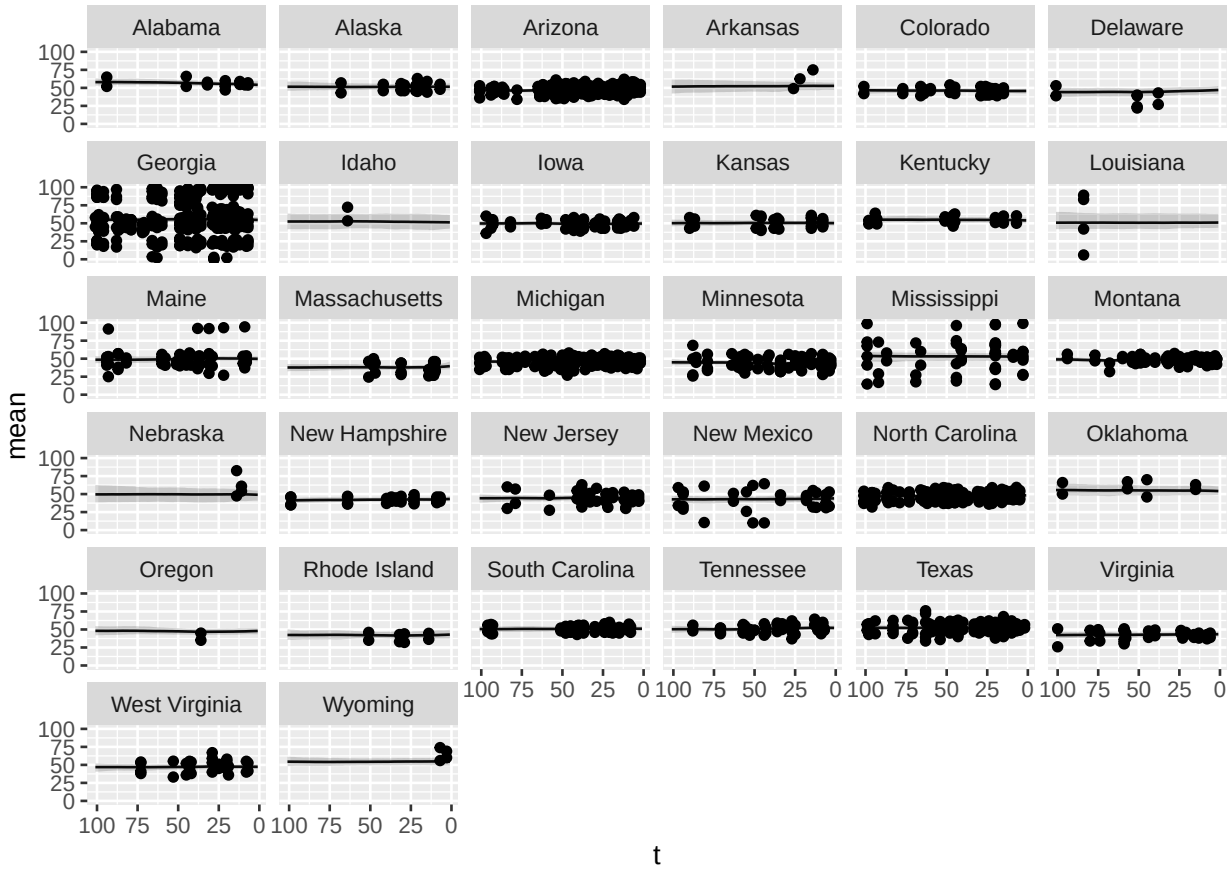


Table 6: Swing State Share Percentage Interval Estimate (Biden)

	Official Model			Informative Prior			Imputation Perturbation		
	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
AZ	48.73	50.44	52.32	48.73	50.44	52.32	48.73	50.44	52.32
NC	48.37	50.09	51.70	48.37	50.09	51.70	48.37	50.09	51.70
MI	49.29	51.50	54.63	49.29	51.50	54.63	49.29	51.50	54.63
WI	49.16	51.29	53.97	49.16	51.29	53.97	49.16	51.29	53.97
MN	48.97	51.18	54.47	48.97	51.18	54.47	48.97	51.18	54.47
TX	47.44	49.31	51.12	47.44	49.31	51.12	47.44	49.31	51.12
FL	48.29	50.02	51.80	48.29	50.02	51.80	48.29	50.02	51.80
PA	48.80	50.74	53.37	48.80	50.74	53.37	48.80	50.74	53.37
GA	47.91	49.88	51.64	47.91	49.88	51.64	47.91	49.88	51.64
OH	47.25	49.59	51.39	47.25	49.59	51.39	47.25	49.59	51.39
NV	48.89	51.07	53.66	48.89	51.07	53.66	48.89	51.07	53.66
NH	48.77	51.16	54.96	48.77	51.16	54.96	48.77	51.16	54.96
IA	47.96	49.85	51.92	47.96	49.85	51.92	47.96	49.85	51.92

```
## Warning: Removed 2716 rows containing missing values (geom_point).
```




```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```

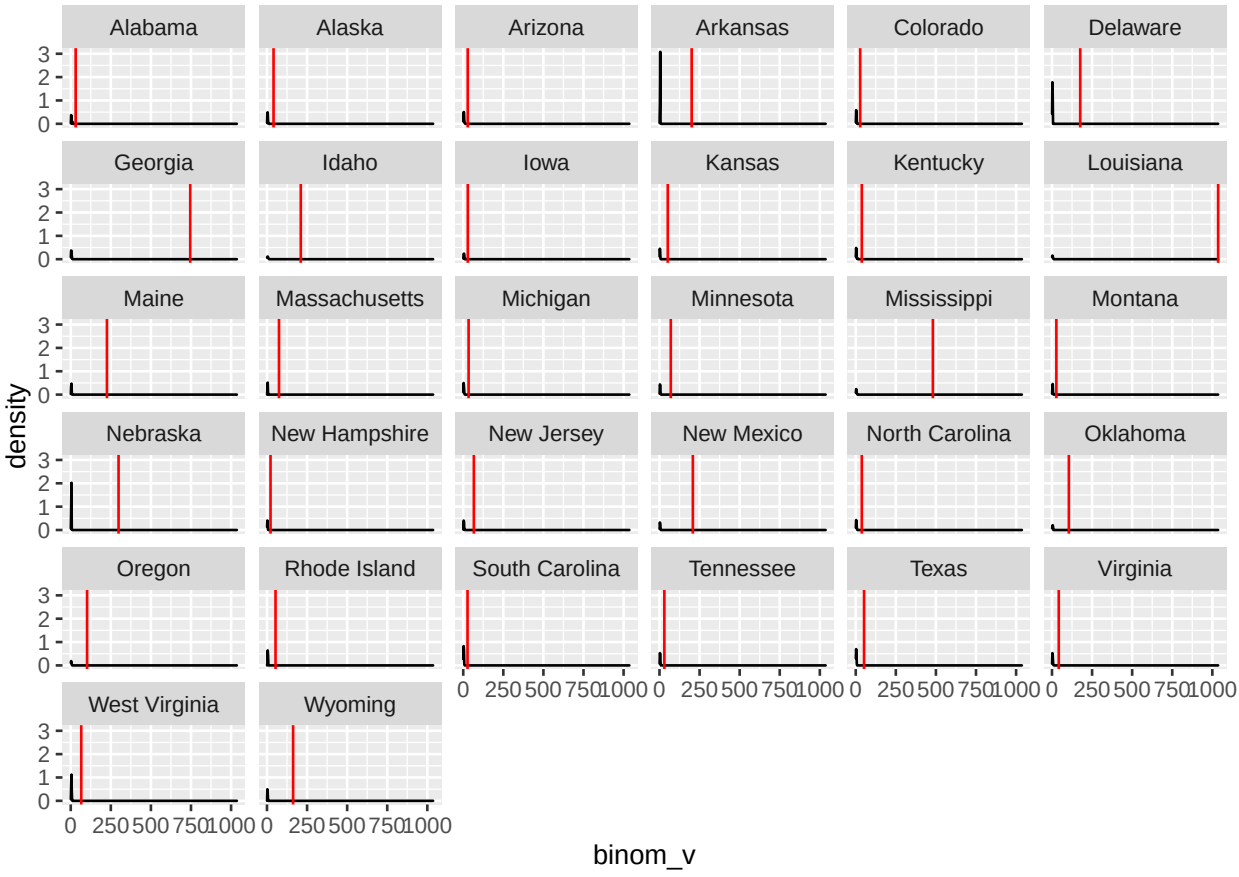
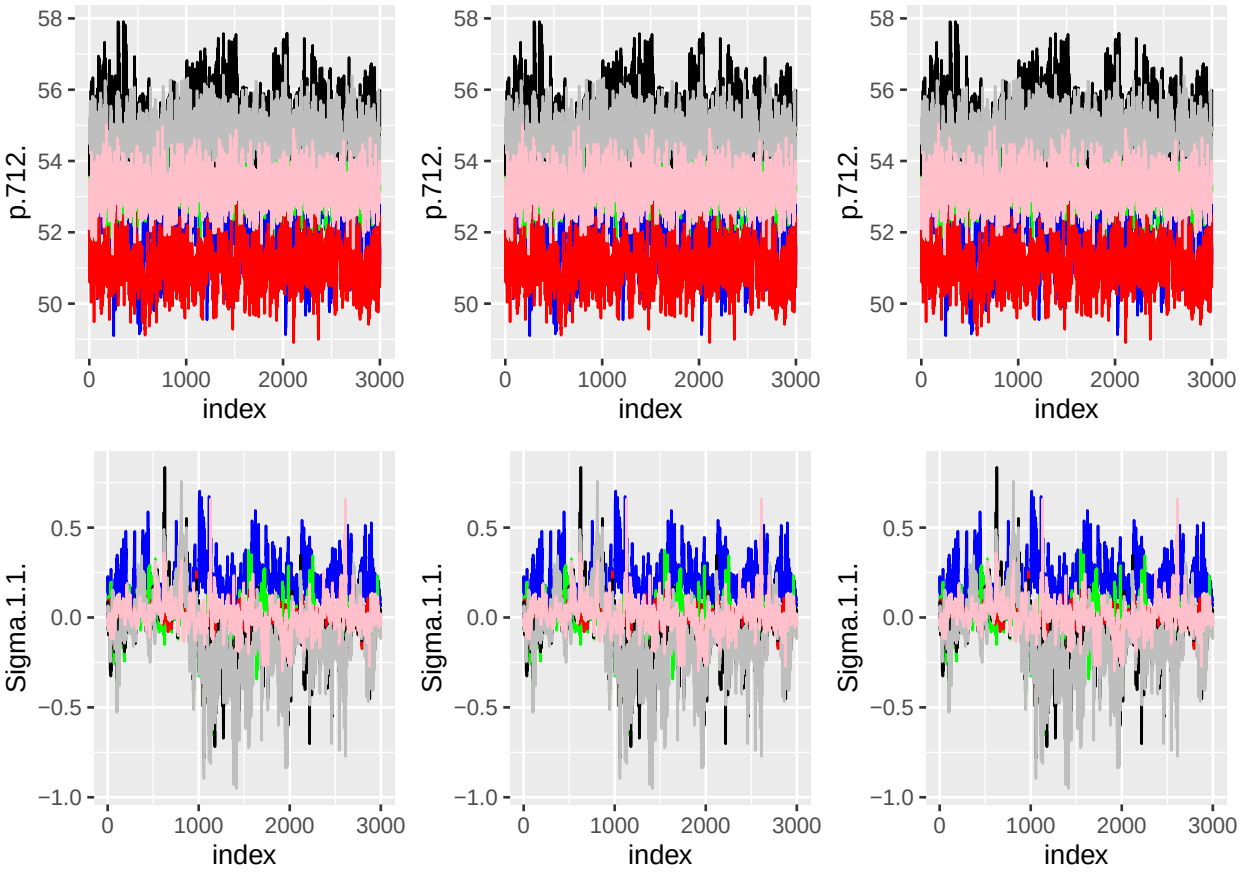

Alabama
Alaska
Arizona
Arkansas
Colorado
Delaware
Georgia
Idaho
Iowa
Kansas
Kentucky
Louisiana
Maine
Massachusetts
Michigan
Minnesota
Mississippi
Montana
Nebraska
New Hampshire
New Jersey
New Mexico
North Carolina
Oklahoma
Oregon
Rhode Island
South Carolina
Tennessee
Texas
Virginia
West Virginia
Wyoming
density
0
1
2
3
0
250
500
750
1000
binom_v

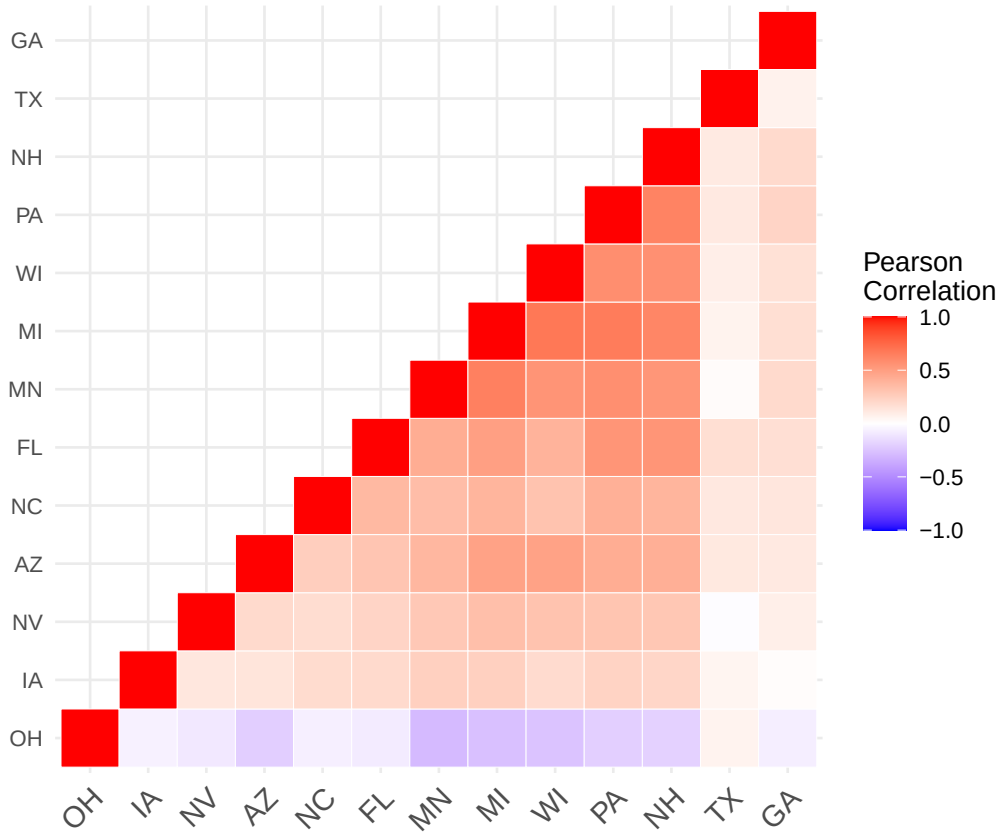
Table 7: Senator All State Winning Probability (REP)

	Official Model	Informative Prior	Imputation Perturbation
Michigan	0.0577	0.0397	0.0690
Minnesota	0.0467	0.0073	0.0513
Arizona	0.8143	0.8043	0.8683
North Carolina	0.5780	0.5623	0.6040
Iowa	0.7930	0.9060	0.8613
Virginia	0.0043	0.0040	0.0200
Georgia	1.0000	0.9993	0.9980
Texas	0.9970	0.9940	0.9960
Alaska	0.9150	0.9263	0.9623
New Hampshire	0.0033	0.0000	0.0037
Alabama	0.9953	1.0000	0.9883
Kentucky	0.9993	1.0000	0.9997
Wyoming	0.9887	0.7537	0.9420
Montana	0.5300	0.4943	0.6280
Kansas	0.8897	0.8627	0.9120
Maine	0.8650	0.8497	0.7253
South Carolina	0.9550	0.9750	0.9657
Mississippi	0.9150	0.9947	0.9507
New Jersey	0.0387	0.0297	0.0320
Tennessee	0.9940	0.9870	0.9720
Nebraska	0.6857	0.5463	0.4890
Massachusetts	0.0000	0.0000	0.0063
Arkansas	0.9353	0.8677	0.8507
Oklahoma	0.9983	0.9057	0.9610
Colorado	0.0910	0.1747	0.2160
New Mexico	0.0327	0.3550	0.1250
West Virginia	0.3583	0.5640	0.5983
Oregon	0.5097	0.3760	0.6117
Delaware	0.4080	0.0903	0.3253
Idaho	0.7337	0.5697	0.4727
Louisiana	0.4960	0.5467	0.5997
Rhode Island	0.0707	0.0460	0.1300

Table 8: Senator All State Vote Share Percentage Interval Estimate (REP

	Official Model			Informative Prior			Imputation Perturbation		
	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
Michigan	43.75	46.11	48.54	43.72	45.98	48.23	44.06	46.32	48.56
Minnesota	41.23	44.35	48.71	41.83	44.24	47.15	41.52	44.55	48.60
Arizona	46.83	48.93	51.01	47.04	48.83	51.00	47.15	49.08	51.18
North Carolina	45.65	48.27	50.71	45.88	48.19	50.74	45.93	48.32	51.01
Iowa	46.35	49.66	53.17	47.02	49.71	52.33	46.76	49.76	52.46
Virginia	40.22	43.48	47.01	39.51	42.67	46.37	40.48	44.38	47.82
Georgia	50.60	54.55	58.68	50.80	55.18	59.48	49.72	53.59	56.93
Texas	49.51	52.33	54.81	48.74	51.99	55.10	49.13	52.11	54.62
Alaska	46.40	51.68	56.20	47.30	50.63	55.10	47.62	51.29	54.42
New Hampshire	40.28	43.03	47.03	39.93	43.02	46.26	40.31	43.27	46.78
Alabama	49.58	54.03	59.48	50.68	54.95	58.34	48.75	53.58	56.73
Kentucky	49.69	54.14	57.92	49.98	53.89	57.04	49.56	52.93	57.23
Wyoming	48.88	54.38	59.94	45.33	50.57	62.02	47.28	51.40	57.44
Montana	45.74	48.09	50.56	45.68	47.99	50.23	46.12	48.38	50.76
Kansas	46.65	50.13	53.86	46.40	50.36	54.54	46.76	50.32	53.70
Maine	46.30	49.84	53.28	45.14	50.91	58.11	44.77	49.76	54.70
South Carolina	47.67	51.17	54.32	48.00	50.79	54.36	47.85	50.78	53.63
Mississippi	46.58	52.72	57.82	49.07	53.69	59.03	47.21	52.05	57.59
New Jersey	41.43	44.51	48.44	40.51	44.49	48.15	42.12	44.64	48.27
Tennessee	48.90	52.18	55.46	48.44	52.28	54.89	47.87	51.68	54.67
Nebraska	40.74	49.62	54.79	43.87	48.31	55.77	42.43	47.88	57.26
Massachusetts	35.33	39.09	46.01	34.64	39.56	44.54	35.24	41.11	46.87
Arkansas	47.00	53.08	57.24	46.28	50.80	56.40	46.03	50.43	57.14
Oklahoma	48.73	54.69	60.64	46.11	54.57	60.37	47.64	52.54	56.24
Colorado	42.11	45.86	49.04	42.85	46.52	49.51	42.68	46.36	49.71
New Mexico	40.01	44.16	48.19	41.88	46.86	51.59	40.62	45.20	49.61
West Virginia	43.89	47.42	50.83	43.69	48.41	52.37	43.89	48.53	52.61
Oregon	43.24	48.04	51.72	40.83	46.74	55.17	42.42	49.17	54.09
Delaware	41.19	47.35	52.80	38.46	44.38	49.21	42.31	46.94	50.82
Idaho	42.58	50.87	61.65	42.04	48.58	55.44	41.83	47.73	58.65
Louisiana	43.76	47.97	63.16	42.80	48.43	60.13	44.08	48.72	52.81
Rhode Island	37.33	42.96	49.73	36.69	44.26	48.34	40.12	44.88	50.19


p.712.
58
56
54
52
50
0
1000
2000
3000
index
Sigma.1.1.
0.5
0.0
−0.5
−1.0




```
mean(as.data.frame(jags_sims_mv$BUGSoutput$summary)$Rhat)
```

```
## [1] 1.035
```

```
mean(as.data.frame(jags_sims_senator$BUGSoutput$summary)$Rhat)
```

```
## [1] 1.338
```

```
mean(as.data.frame(jags_sims_house$BUGSoutput$summary)$Rhat)
```

```
## [1] 2.457
```