UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

A Bayesian Approach for Combining Stratified Election Polls

Samuel Zalewski

Universitat Politècnica de Catalunya

June 7, 2024

Contents

1 Introduction 2
1.1 Stratified Polling 2
1.2 Bayesian Model Averaging 3
1.3 Challenges of USA Election Polls 4

2 A Bayesian Model for Interpreting Polls 5

3 Data Collection 7
3.1 Polls 7
3.2 Voter Turnout Rate 8
3.3 Population 11

4 Model Fitting 11

5 Results 16

6 Discussion 22

References 24

1 Introduction

1.1 Stratified Polling

As seemingly instant press coverage and media transmission becomes more and more expected, up-to-date election polls are in ever higher demand. Accurate election polls are essential for a variety of reasons: for newscasters and journalists to provide accurate information, political strategists to better understand public perception, and even for ordinary citizens to better understand the context of their own decisions in coming elections. However, representing the voting intentions of an entire nation, or even a national subdivision, can quickly become an expensive task when considering the sample sizes required for adequate statistical power.

For this reason, stratified sampling has become a tool of foundational importance in modern polling to reduce bias in estimates while cutting costs. Stratified sampling is, in its essence, the sampling of distinct demographic groups, and then weighing the results of each demographic group according to their representation at the political division being estimated. In *Understanding Elections through Statistics*,[1] stratified sampling is formulated as the following:

$$X_s = \sum_{g=1}^{G} w_g X_g$$

with G being the number of groups, or strata that the population is divided into. X_g is the mean response of a given stratum g, and then w_g is the associated weight for that stratum. This weight will be, at its simplest, the population of this strata, but it can further complicated by multiplying the weight by things such as the actual voter turnout, or in some way adjusting for the difference in polling vs voting.

The benefits of stratified polling are clear: by recognizing the difference in voting behaviour in different demographics, you can improve the accuracy of polls while still cutting back on total sample size. However, there are some challenges that come with stratification, one of the main ones being the increasing difficulty of stratifying multiple demographics at once. For example, say you would like to stratify by n age groups in m different states. In order to accurately arrive at an estimator X_g, you will need accurate population and polling data for every combination of n and m. For just two groups of stratification, the required n by m matrix of information is often obtainable. But if you wanted to stratify by, say, age, state, and then gender, the dimensionality of the necessary polling and population matrices would be increased to three, making the cells of each cross section much more difficult to find information on. With further stratification, finding available data becomes nearly impossible. This is due not just to the price of scarce information, but also because privacy concerns prevent agencies from revealing information that could be used to trace the

[1] Ole J. Forsberg, *Understanding Elections through Statistics* (Boca Raton, Florida: CRC Press, 2020).

identities of too small a group of individuals. The fundamental problem becomes, although there are many different ways to demographically stratify in election polls, there isn't necessarily a clear way to reconcile the information obtained from these distinct stratification methods.

1.2 Bayesian Model Averaging

To solve the challenge of combining multiple models with varying levels of credibility, several approaches have been developed, with more thoroughly Bayesian methodologies being developed in the second half of the 20th century. *Bayesian Model Averaging: A Tutorial*[2] provides a straightforward summary of this Bayesian strategy. If there is some quantity of interest δ, its posterior distribution can be estimated by the following equation:

$$\mathrm{pr}(\delta \mid D) = \sum_{i=1}^{I} \mathrm{pr}(\delta \mid M_i, D)\mathrm{pr}(M_i \mid D).$$

This is simply the sum of each the posterior distribution for each model M in set I, multiplied by the posterior's probability $\mathrm{pr}(M_i \mid D)$. As a weighted sum, the probability of each posterior being considered is normalized relative to the other models used in the calculation of δ, formulated by:

$$\mathrm{pr}(M_i \mid D) = \frac{\mathrm{pr}(D \mid M_i)\mathrm{pr}(M_i)}{\sum_{i=1}^{I} \mathrm{pr}(D \mid M_i)\mathrm{pr}(M_i)}.$$

Here, $\mathrm{pr}(D \mid M)$ is the likelihood of the data given some model M, and $\mathrm{pr}(M)$ is the prior probability of a model. By normalizing the model probabilities $\mathrm{pr}(M \mid D)$ with the sum of all models' probabilities, we ensure that the scale of the estimated quantity δ is correct, while prioritizing the weight of higher-likelihood models.

While BMA sounds promising, and can certainly be a powerful tool in statistical analysis, in practice it often brings about many challenges. In BMA, priors are used in several steps, one of the most imporant being the comptation of marginal likelihood of the data D for each model M.

$$\mathrm{pr}(D \mid M_i) = \int \mathrm{pr}(D \mid \theta_i, M_i)\mathrm{pr}(\theta_i \mid M_i)\, d\theta_i$$

This marginal likelihood of a given a model M_i is computed by integrating over all values of the space of its parameters θ_i. This integral can often be computationally intensive, but as been made more feasible in recent years with the development of several packages that will be discussed in later sections.

[2] Jennifer A. Hoeting et al., "Bayesian Model Averaging: A Tutorial," *Statistical Science* 14, no. 4 (1999): 382–417.

Improper selection of priors and other model assumptions can also have disastrous effects on the model weights, as overly tight priors can effectively drive a model's marginal likelihood to zero, excluding it entirely from the summation of $\mathrm{pr}(\delta \mid D)$.

Given the problem specified in 1.1, I propose a method of combining the results of multiple polls, with each model M representing a different way of stratifying a voting population (by age, gender, etc.). With BMA, we can combine the information from all available stratifications without the need for multidimensional matrices for all possible demographic cross-sections.

1.3 Challenges of USA Election Polls

Before continuing on to a formal model declaration, there are a few important notes to address about the country to which the model will be applied. In the United States (USA), the winner of presidential elections is determined by a system known as an electoral college. In this system, every state effectively has its own presidential election, and the winner of each state receives a predetermined amount of points for winning this state. Instead of being chosen by the popular vote, the winner is decided by whichever candidate receives more points from the individual states.

In recent history, there have been two elections in which the winner had lost the popular vote but won the election due to winning more points from the electoral college. The first was 2000, where Republican George W. Bush, with 47.9% of the popular vote and 271 electoral points won against Democrat Al Gore, who had 48.4% of the popular vote and 266 votes. Later, in 2016, Democrat Hillary Clinton won the popular vote at 48.2% versus Republican Donald Trump's 46.1%, but lost because Trump had secured 304 electoral points, while Clinton had won only 227.

Unsurprisingly, this electoral college brings about significant difficulties in predicting election outcomes. Although national elections are decided by the results of individual states' elections, the amount of points granted by each state aren't necessarily proportional to the population. This is because the points for a given state calculated with a baseline two, and then adding more points depending on the distribution of the national population. This generally results in granting more voting power to less populous states. For example, Wyoming with three electoral votes has an effective vote of 3.19 per capita ((3 electoral votes/pop. of Wyoming 581,381)*(national population 333.3 million/538 total votes)), whereas Texas has an effective vote per capita of closer to 0.85.

Because the majority winner of the each state is all that matters in the presidential election, regardless the margin won by, the outcome of noncompetitive states are often of little interest. For this reason, electoral discourse usually pivots around a small set of key states with high uncertainty, dubbed "swing states". These few states will most likely determine the winner of the most electoral points, and the presidential election. These states will be the main focus of a Bayesian model.

2 A Bayesian Model for Interpreting Polls

Given the following terms:

- S: Number of states
- D: Number of demographic divisons
- K: Number of voting options
- p_{sd}: Total population of demographic d in state s

The estimated total votes for a candidate k in a state s can be estimated by the formula:

$$V_{sk} = \sum_{d \in D} \phi_{sd} \cdot \theta_{sdk} \cdot p_{sd}$$

Where ϕ_{sd} is the turnout rate for demographic d in state s, and θ_{sd} is the proportion of demographic d in state s voting for candidate k. Note that this is simply a variation of the equation stated in 1.1, but now it it multiplies the population of a given demographic group by the voter turnout for that group, giving a more accurate estimate of the total votes V_{sk} that will be received. The behaviour of these model parameters can be described by the following set of equations:

Model Components

$$\begin{aligned}
\text{Voting intentions:} \quad & \theta_{sdk} \sim \text{Dirichlet}(\alpha_{sd}) \quad \forall s \in S, d \in D \\
\text{Alpha parameters:} \quad & \alpha_{sd} = (\alpha_{1sd}, \ldots, \alpha_{Ksd}) \quad \forall s \in S, d \in D \\
\text{Logit of turnout:} \quad & \psi_{sd} = \text{logit}(\beta + \gamma_s + \delta_d) \quad \forall s \in S, d \in D \\
\text{Adjusted turnout mean:} \quad & \mu_{sd} = \text{inv logit}(\psi_{sd}) \quad \forall s \in S, d \in D \\
\text{Adjusted turnout:} \quad & \phi_{sd} = \text{Normal}(\mu_{sd}, 0.1) \quad \forall s \in S, d \in D
\end{aligned}$$

Priors

$$\begin{aligned}
\text{Demographic effect:} \quad & \delta_d \sim \text{Normal}(0, 1) \quad \forall d \in D \\
\text{State effect:} \quad & \gamma_s \sim \text{Normal}(0, 1) \quad \forall s \in S \\
\text{2024 turnout rate:} \quad & \beta \sim \text{Beta}(100, 65)
\end{aligned}$$

In order to obtain parameters of voting intention θ, we can use the raw numbers from polls as the input of a Dirichlet distribution, which will return a random value θ according to its probability given the vector α, which contain the total number of survey responses for each respective candidate. The probability density of a Dirichlet is given by:

$$f(\theta_1, ...\theta_K, \alpha_1, ...\alpha_K) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_i - 1}$$

$$\text{where } B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$$

$$\text{and } \alpha_0 = \sum_{i=1}^{K} \alpha_i$$

Simply put, the Dirichlet distribution gives us the density for some mutually exclusive multinomial probabilities θ which add up to 1. In the context of election polling, there will be a set of K parameters θ generated, each for a different voting option. With a vector of poll responses for a given state and demographic combination α_{sd}, the Dirichlet will generate a parameter θ_{sdk} to estimate the support for each candidate k among that state and demographic combination. In Bayesian statistics, the Dirichlet is typically used as the conjugate prior for the multinomial distribution. When dealing with smaller populations, it might be necessary to simulate the results of the a multinomial distribution given these values of θ, however due to the law of large numbers we can expect the behaviour of states with population in the millions to converge with the values of θ, so in this case further simulation isn't required.

Next, turnout rates are modelled using a logistic regression model, ensuring the they're confined to $[0, 1]$. The logit of this turnout mean ψ_{sd} is calculated in a hierarchical form, starting off with some national turnout rate β and then adding the effects that both a demographic δ_d and state γ_s have on turnout. This mean is then converted from logistic format ψ_{sd} to an applicable percentage μ_{sd}. The adjusted turnout is then assumed to follow a normal distribution with this μ_{sd} as the average, and a standard deviation of 10%. This is a very generous assumption, that the real voter turnout should follow a normal distribution with σ of 10%, but because there is only one sample for each combination of state and demographic, uncertainty is artificially raised so that the marginal likelihood stays manageable.

The prior distributions for the demographic and state effects have both been set to Normal(0,1), indicating low-information priors. For the turnout rate of the 2024 election, I've selected as Beta(100,65) which gives a mean of 60.6% and a standard deviation of 3.8%. This is just a general guess of what the voter turnout may look like, with a pretty wide range on either side. Although there are some methods developed to predict national turnout turnout rate using information from historical trends and recent polls, that lies outside the scope of this paper, and we will assume that the 2024 national turnout will be reasonably within this range.

3 Data Collection

In order to fit my model and generate estimates for future elections, there are three important data I need: poll responses, voter turnout rates, and population. For each type of model that I create, I need to know these three data points for each state and demographic group.

As of June 2024, most US journalists are considering there to be seven 'swings states' - that is, seven crucial states that will decide the outcome of the election.[3] These states are Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. Because of their relevance, there is more poll data available for them, and I will be selecting these seven as the states to include in my model. For demographic stratifications, I will be making three seperate models to compare: one by gender, by race, and then by age.

3.1 Polls

When searching for poll data to use, I wanted to have to some consistency across all states and demographic stratifications. Unfortunately there aren't many polling companies that published data for all seven swing states for the same time period. For this reason, I ended up using the *Morning Consult Swing States Tracking Poll*, conducted from April 8-15 in 2024.[4] Using the same polling company for all my data has the benefit of reducing variance in my model, but there is the potential that this has the tradeoff of increased bias. While many pollsters do publish the raw data from their survey respondents, they typically also publish their estimated overall estimates for entire states or nation. The difference in weights that each polling company has to create these estimates is known as the "house effect". That is to say, given a set of responses, each polling company will likely result in different final predictions.

By using the raw responses from the polls instead of their estimated percentages, there are two benefits: 1) the house effect, which is a significant source of bias in polling, can be eliminated, and 2) the raw data from the polls can be used as input for the Dirichlet distribution, which will then generate θ values that accurately reflect the degree of certainty given our sample sizes.

The Morning Consult poll data is formatted as the following: each state has its own page, and then within that page you can see both the percentages and raw numbers for each candidate, sorted by demographic. For example, here are the first 7 rows from the 'Arizona' poll. For the purpose of this paper, only four possible voting options will be considered: Biden, Trump, RFK Jr., and other. Although

[3] U.S. News & World Report, "7 Swing States That Could Decide the 2024 Presidential Election," Accessed June 1, 2024, 2023, accessed June 1, 2024, https://www.usnews.com/news/elections/articles/7-swing-states-that-could-decide-the-2024-presidential-election.

[4] Morning Consult, "Swing States Tracking Poll #2404025: April 08-15, 2024," Accessed June 1, 2024, 2024, accessed June 1, 2024, https://pro-assets.morningconsult.com/wp-uploads/2024/04/Bloomberg_2024-Election-Tracking-Wave-7.pdf.

there are other third-party candidates besides RFK Jr, their polling numbers are not significant enough to warrant including in the model, so they were aggregated into the 'Other' column. Another detail of importance is the fact that 'would not vote' was an option on the poll, but I will be including it in the 'Other' category. Because the poll is directed at registered voters, the percentages of 'would not vote' were generally very small, in the low single digits. The following table is contains the first seven rows from the Arizona poll. There are many more demographic rows in the original report, including categories like religion, employment status, and vaccination status, but as stated earlier, I will just be using the demographic data for gender, age, and race.

Demographic	**Biden**	**Trump**	**RFK Jr.**	**Other**	**Total N**
Registered Voters	40% (318)	46% (366)	7% (54)	7% (45)	801
Gender: Male	40% (146)	51% (186)	6% (21)	3% (11)	368
Gender: Female	40% (172)	42% (180)	8% (32)	11% (47)	433
Age: 18-34	35% (71)	39% (80)	12% (24)	10% (21)	205
Age: 35-44	46% (54)	40% (47)	4% (5)	9% (11)	117
Age: 45-64	38% (102)	52% (139)	5% (12)	6% (16)	269
Age: 65+	43% (91)	47% (100)	6% (12)	4% (8)	211

Table 1: 'Morning Consult' Arizona Poll Results

3.2 Voter Turnout Rate

Finding voter turnout was mostly a straightforward process. Every four years since 2000, the US Census Bureau publishes demographic statistics for that year's election. These tables include gender, race, and age categories.

Sex, Race, and Hispanic-Origin	**Total Population**	**Percent Voted (Total)**
Total	5,638	64.7
Male	2,739	60.4
Female	2,899	68.9
White alone	4,840	65.1
White non-Hispanic alone	3,140	76.0
Black alone	279	63.9
Asian alone	206	52.0
Hispanic (of any race)	1,800	45.2
White alone or in combination	4,966	65.3
Black alone or in combination	344	68.3
Asian alone or in combination	226	56.2

Table 2: US Census Voter Data, Arizona 2000

Not shown in the above table is a column of confidence intervals for estimated voter turnout. At the national level, these confidence intervals are low, nearing just a few percentage points. For individual states' demographic groups, on the other hand, these confidence intervals can become quite large. For example, the estimated turnout for Asian American and Pacific Islander voters in the 2020 election was 36.3 ± 12.8% for a 90% confidence interval. Given the strictly Bayesian nature of this paper, these confidence intervals won't be used as direct computational input, but the possible wide range of their true values will be acknowledged in decisions involving the variance of turnout estimations.

One problem that did arise with the census data is the age buckets used have changed in the past 24 years. In 2000, 2004, and 2008, age groups were bucketed into five groups: 18-24, 25-44, 45-64, 65-74, and 75+. However, starting in 2012 they are formatted as 18-24, 25-34, 45-44, 45, 64, and 65+. As seen in Table 1, Morning Consult separates its respondents into four categories: 18-34, 35-44, 45-64, and 65+. Fortunately, these categories coincide in a way that we can neatly organize the data into three age groups: 18-44, 45-64, and 65+. This will require some row aggregation, but because we have both the percentages and total populations for each group in the census data, we can easily calculate the turnout rate for aggregated age groups.

Looking at the historical voter turnout data for different demographic groups, we can already see which types of stratification may lend themselves to having easier to predict patterns. Gender/state turnout rates follow a consistent relationship with one another, age/state rates have a hierarchy that generally holds for each year, and race/state rates are the most chaotic, with only a general rule of thumb relating the different turnout rates.

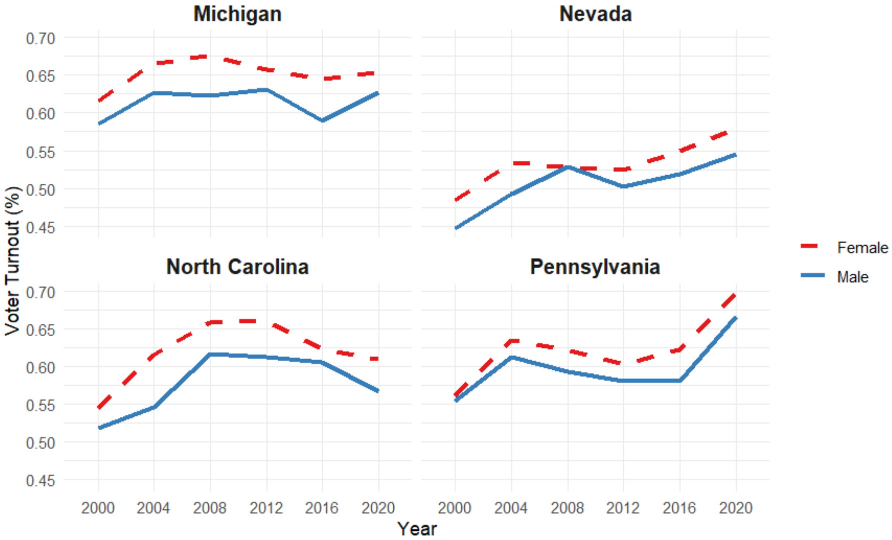


Figure 1: Turnout Rates per year by Gender

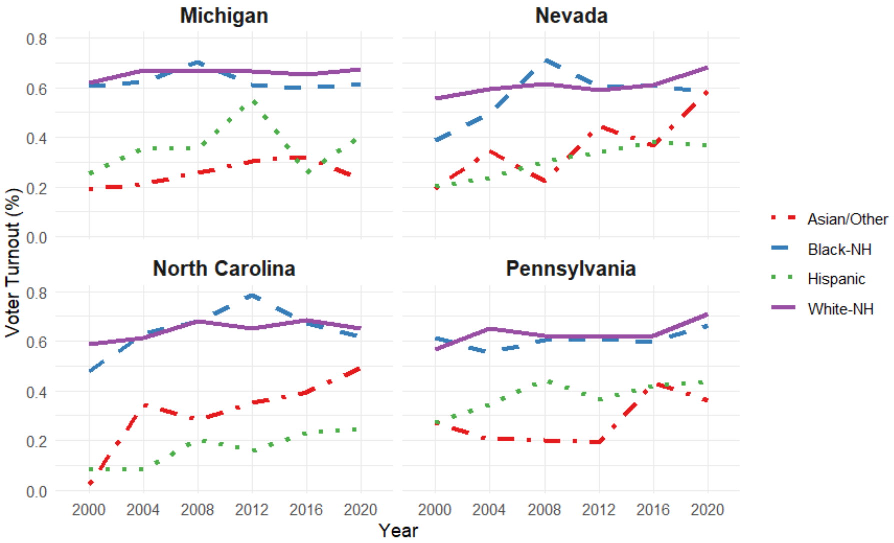


Figure 2: Turnout Rates per year by Race

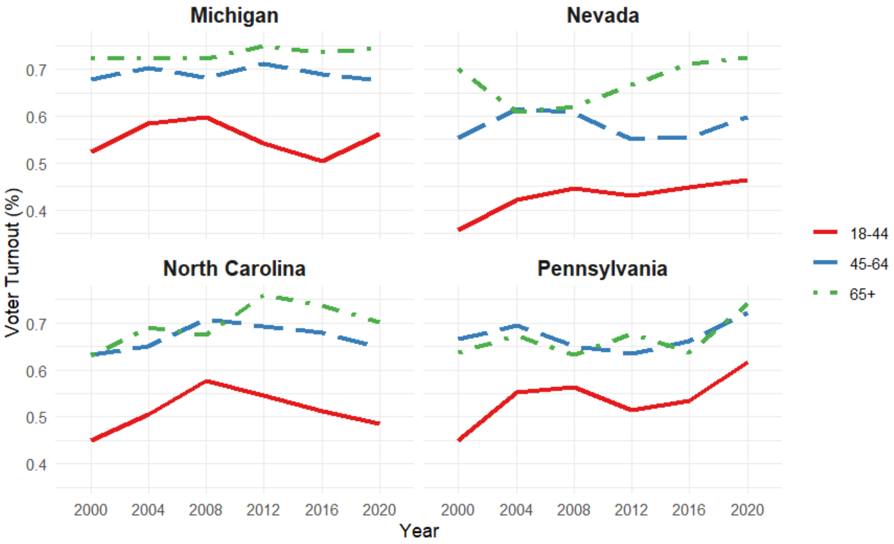


Figure 3: Turnout Rates per year by Age

3.3 Population

Because the US census data also includes the total population for each demographic group in each state, we can just use them directly as the population size in our model. A possible issue is that the census data is from 2020, but although the population has increased since then, the proportions are likely close enough. Although the census counts by race, age, and gender were all theoretically sampling the same population at the same time, there are some some slight differences in the total sum. The total sum of voting age population in the seven swing states is counted as 47870000, 45435000, and 46423000, for gender, race, and age census, respectively. To adjust for the difference, I simply multiplied the total predicted votes in the race and age model by the factor by which the population differs from the total of the gender model.

4 Model Fitting

With the data now established, the model is ready to be executed. To compute the model described in section 2, I wrote it in STAN with the rstan package for R 4.3.2. The MCMC simulations were ran across 4000 iterations with 4 chains, with the resulting trace plots indicating succesfull convergence of parameters:

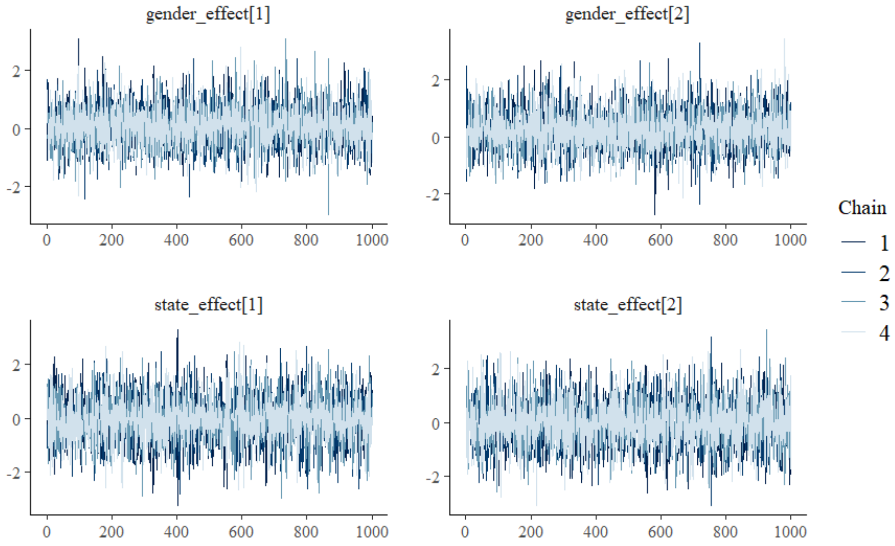


Figure 4: Age/State Model Trace Plots

These trace plots were checked across all three models to verify convergence. Next, we can view the estimated coeffecients of our model. These coeffecients will also include values such as their standard deviation, quartile values, and Rhat values. Rhat values near 1 are good sign that values have converged correctly.

Table 3: Summary of Gender, Race, and Age Model Effects

Effect	Mean	SE	SD	2.5%	25%	50%	75%	97.5%	N_eff	Rhat
gender_effect[1]	-0.02	0.01	0.34	-0.70	-0.24	-0.01	0.22	0.61	1113	1.01
gender_effect[2]	0.14	0.01	0.34	-0.54	-0.08	0.15	0.38	0.77	1100	1.01
state_effect[1]	-0.19	0.01	0.35	-0.85	-0.43	-0.20	0.04	0.52	1172	1.01
state_effect[2]	-0.13	0.01	0.35	-0.78	-0.37	-0.14	0.10	0.58	1192	1.01
state_effect[3]	0.19	0.01	0.35	-0.48	-0.06	0.18	0.43	0.89	1196	1.01
state_effect[4]	-0.28	0.01	0.35	-0.94	-0.53	-0.29	-0.05	0.41	1177	1.01
state_effect[5]	0.04	0.01	0.35	-0.61	-0.21	0.03	0.27	0.73	1169	1.01
state_effect[6]	0.09	0.01	0.35	-0.58	-0.16	0.08	0.33	0.80	1209	1.01
state_effect[7]	0.50	0.01	0.36	-0.18	0.25	0.49	0.74	1.22	1196	1.01

Effect	Mean	SE	SD	2.5%	25%	50%	75%	97.5%	N_eff	Rhat
state_effect[1]	-0.22	0.01	0.32	-0.85	-0.43	-0.22	-0.01	0.41	734	1
state_effect[2]	-0.45	0.01	0.32	-1.09	-0.66	-0.45	-0.25	0.18	728	1
state_effect[3]	-0.10	0.01	0.32	-0.74	-0.31	-0.10	0.11	0.52	722	1
state_effect[4]	-0.18	0.01	0.32	-0.81	-0.39	-0.18	0.03	0.45	742	1
state_effect[5]	-0.23	0.01	0.32	-0.86	-0.44	-0.22	-0.02	0.41	737	1
state_effect[6]	-0.11	0.01	0.32	-0.74	-0.32	-0.11	0.10	0.52	723	1
state_effect[7]	0.06	0.01	0.32	-0.56	-0.14	0.07	0.27	0.70	721	1
race_effect[1]	-0.96	0.01	0.32	-1.57	-1.16	-0.96	-0.75	-0.32	719	1
race_effect[2]	0.20	0.01	0.31	-0.43	0.00	0.19	0.40	0.82	709	1
race_effect[3]	-0.96	0.01	0.32	-1.59	-1.16	-0.96	-0.75	-0.34	713	1
race_effect[4]	0.50	0.01	0.31	-0.12	0.29	0.49	0.70	1.12	705	1

Effect	Mean	SE	SD	2.5%	25%	50%	75%	97.5%	N_eff	Rhat
state_effect[1]	-0.15	0.01	0.32	-0.76	-0.37	-0.16	0.06	0.50	926	1
state_effect[2]	-0.08	0.01	0.33	-0.70	-0.31	-0.09	0.14	0.57	943	1
state_effect[3]	0.21	0.01	0.33	-0.41	-0.02	0.20	0.43	0.85	956	1
state_effect[4]	-0.23	0.01	0.32	-0.85	-0.45	-0.23	-0.01	0.41	931	1
state_effect[5]	0.06	0.01	0.32	-0.55	-0.17	0.05	0.29	0.71	935	1
state_effect[6]	0.06	0.01	0.33	-0.57	-0.16	0.06	0.28	0.72	942	1
state_effect[7]	0.53	0.01	0.33	-0.09	0.31	0.53	0.76	1.19	959	1
age_effect[1]	-0.32	0.01	0.31	-0.95	-0.53	-0.31	-0.10	0.27	892	1
age_effect[2]	0.32	0.01	0.32	-0.31	0.11	0.32	0.54	0.92	893	1
age_effect[3]	0.53	0.01	0.32	-0.11	0.32	0.54	0.75	1.13	901	1

At a first glance, the models seems to be fitting alright. The gender effects reflect what we know about women voting at higher rates than men, voter turnouts increase with age, and states like Nevada have lower turnouts than states like Wisconsin. High N eff values and Rhats close to 1 also give us confidence about the chain convergence. However, there is a cause for concern: the standard deviation for all the model effects are quite high, all being over 30%.

Recalling back to the model specifications set in 1.2, the turnout rates were expected to follow a normal distribution with a standard deviation of 10%. This already felt generous, but looking at the model summaries it seems that this restriction of 10% SD may be an issue, as it can have quite severe effects on marginal probabilities if not complied.

Calculating marginal probabilities can often be a computationally intensive task, so I will be leveraging a method called bridge sampling, with the *bridgesampling* package for R. Bridge sampling is an advanced method of marginal probability estimation that involves sampling from both a model's posterior and a reference distribution, using some bridge function $h(\theta)$. The details of this bridge function and its relation to marginal probability are quite complicated, but are explained in depth in the *bridgesampling* documentation.[5] With the marginal probabilities calculated, we can view their logarithms, which is the default output:

```
> c(bridge_gender$logml, bridge_race$logml, bridge_age$logml)
[1] -125.1748 -206.0258 -132.9763
```

or their converted values into more intuitive marginal probabilities:

```
> mls
[1] 4.337806e-55 3.342990e-90 1.774618e-58
```

As you can see, these marginal likelihoods are quite small. The next step in model averaging is standardizing these marginal likelihoods as a fraction of the total likelihood of all three models.

```
> mls /sum(mls)
[1] 9.995911e-01 7.703486e-36 4.089378e-04
```

As a result of these extremely small marginal likelihoods, the Bayesian averaging factors are extremely skewed towards the most likely model, the gender model. This effectively makes the BMA model an exact replica of the Gender, with the weights for the other two models being shrunk to essentially zero.

There are a few ways to proceed from here. Firstly, instead of model averaging, we can consider this likelihood comparison as a type of model selection. In this case, the marginal likelihoods tell us that the gender stratification is the most promising of the three, and we would simply discard the other two.

Another option is to investigate the sources of these large standard deviations that are driving down model likelihoods. To do this, we can investigate the residuals from the models, calculating the difference between the turnout rates estimated from the model parameters and the actual turnout rates. Let's look at the residuals for the race model, the one which is considered the worst in terms of likelihood.

[5] Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers, *bridgesampling: An R Package for Estimating Normalizing Constants*, Software manual, Accessed: 2024-06-01 (The Comprehensive R Archive Network, 2017), accessed June 1, 2024, https://cran.r-project.org/web/packages/bridgesampling/vignettes/bridgesampling_paper_extended.pdf.

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Value	-0.48859	-0.09693	-0.02825	-0.02984	0.03541	0.40328

Table 4: Race Model Residual Summary

We can see from the summary table that on average, the model overestimates voter turnout, with the mean residual being -2.9%. The 1st and 3rd quartile values are -9.7% and +3.5%, respectively. A residual plot might help us to observe any important trends:

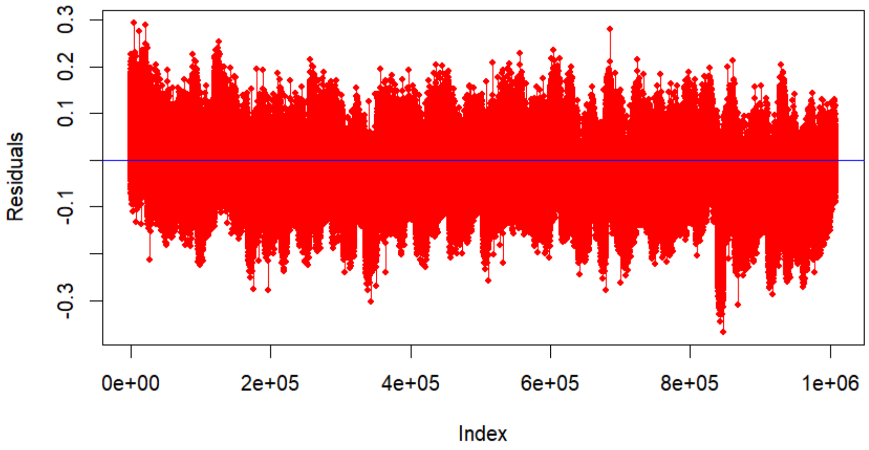


Figure 5: Race Model Residual Plot

Because my data matrix for the model was organized by year of the election, we can view the x-axis of the residual plot as a time component. From this, we are able to see specific years that contained particularly strong residuals. The leftmost values represent 2020, which had a notably high turnout rate, and the rightmost values are from the year 2000, which not just had low for a few ethnic groups, but also contained notoriously large confidence intervals in the US census data.

The slight trend noticeable in the plot doesn't seem strong enough to necessitate a trend component in the model, but it does seem like the relationship between the given national turnout rate specific demographic turnout rates could be better. One adjustment I tried was changing the relationship between national turnout rate and the state/demographic effects from an additive one to a multiplicative one, but this actually yielded slightly worse results.

Given the fact that some of the US census data contains uncertainty, particularly with smaller population subgroups that have less impact on the final election outcome, it may be worth considering increasing the model assumptions to allow a larger standard deviation between predicted and actual voter turnout rates. Increasing the relationship between predicted turnout and actual turnout to the following:

```
turnout_race[i, y] ~ normal(predicted_turnout_race, 0.5);
```

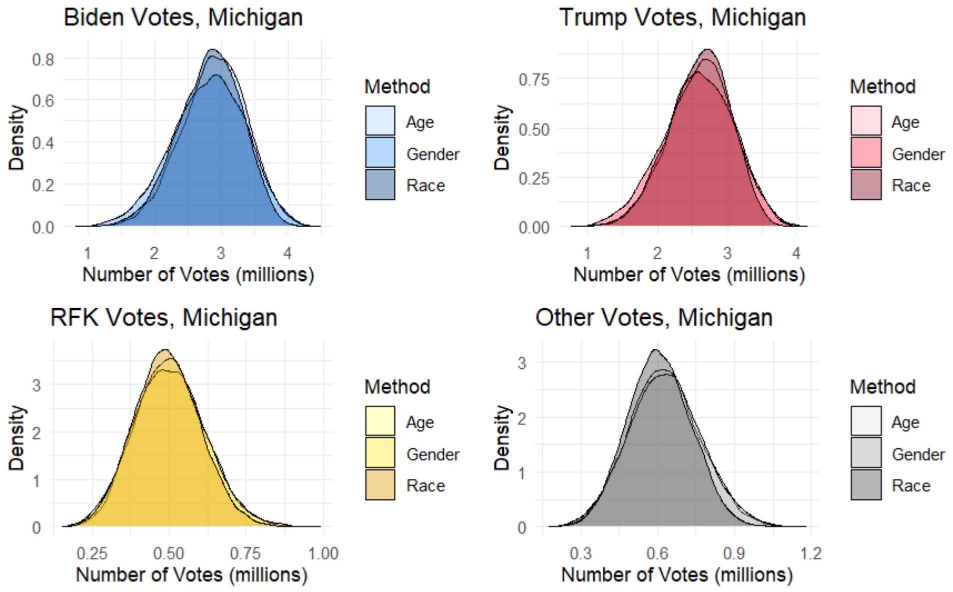
the expected SD for all models' predicted turnout rates will now be 50%, which seems concerningly large, but it's essential to understand how this parameter affects relative posterior probabilities for each model. In the context of BMA, adjusting this value is basically modifying the amount of tolerance we have for incorrect predictions of voter turnout from the census data. By increasing this SD, we are accepting a higher chance of incorrect voter turnout rates in exchange for receiving more information about the poll responses from different demographics, the latter of which does not have an effect on the model likelihood. Here are the new BMA weights for the gender, race, and age models:

```
> posterior_probs
[1] 0.783850155 0.002210672 0.213939173
```

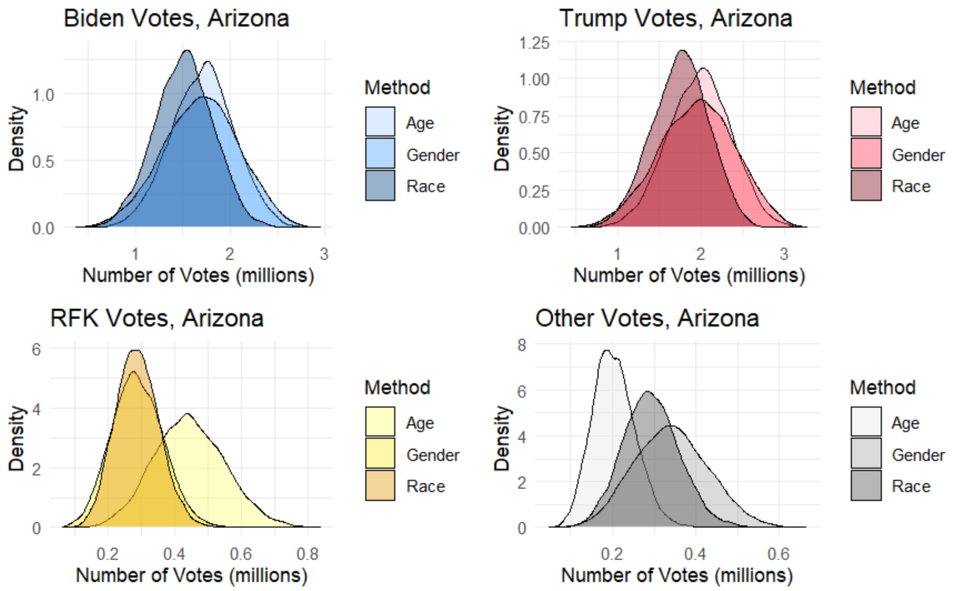
Like before, the gender model is considered the best, but now the likelihoods of other models are large enough that they can be considered in the averaging process. According to these weights, the gender model will make up the bulk of the BMA model, at 78.4%, then the age model is weighted at 21.4%, and the contribution of the race model will be minimal but nonzero, at 0.2%. By multiplying the posterior curves of each of these models by these weights and summing the totals, we can arrive at our new BMA posterior distribution.

5 Results

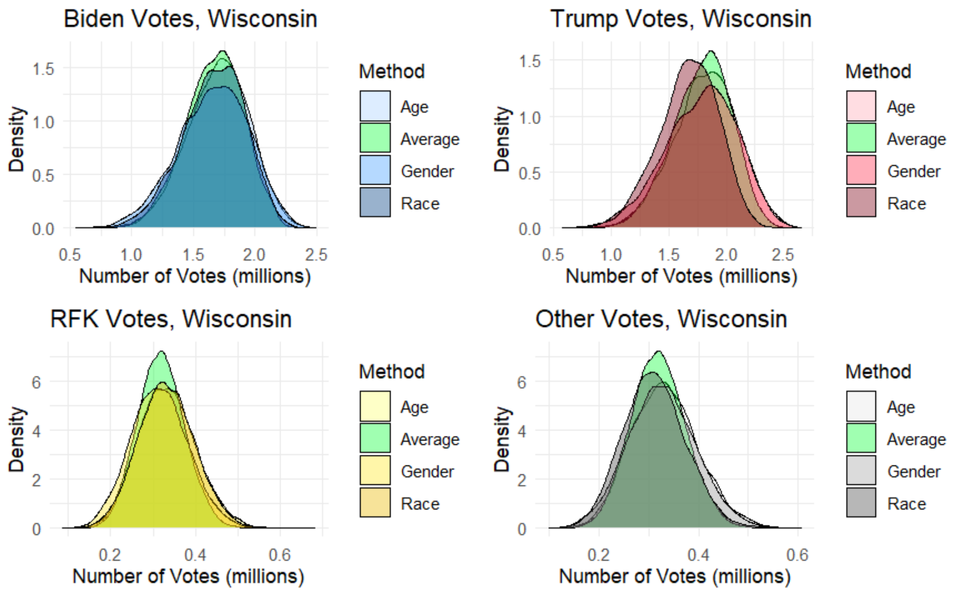
Before seeing the behaviour of the final BMA model, we should first get an understanding of each individual model. Overall, the three models all generated similar posterior distributions for vote counts of the 2024 US election. The following plots show the posterior curves of votes for select states, showing the predicted distribution of votes in the age, gender, and race models. Most states looked similar to the following plots from Michigan, where models had similar peaks, but varying widths of credibility:



Arizona was one of the states with the most disagreement between models. The curves of RFK and Other votes are more separated, but for Biden and Trump, who will both be receiving the majority of votes and are the primary focus of election predictions, the curves align a little better.



The fact that three plots generally coincide is a good sign. This means that instead of averaging out a compromise between competing models, these models can be used to support the claims of the other. Let's take at the posterior curves for Wisconsin's votes, now with the BMA for comparison:



As hoped, the posterior distributions generated by Bayesian Model averaging generate tighter curves with higher peaks, indicating a higher degree of belief around a smaller range of possible votes. For a more precise comparison we can generate 90% credible intervals for each candidate in each model. Let's continue using Wisconsin as an example:

Model	Biden	Trump	RFK	Other
Gender	[1.16, 2.10]	[1.23, 2.26]	[0.21, 0.43]	[0.21, 0.44]
Race	[1.21, 2.05]	[1.20, 2.06]	[0.22, 0.44]	[0.21, 0.42]
Age	[1.27, 2.10]	[1.34, 2.26]	[0.23, 0.45]	[0.23, 0.45]
BMA	[1.26, 2.03]	[1.35, 2.17]	[0.23, 0.41]	[0.24, 0.42]

Table 5: Wisconsin's 90% Credible Intervals for Election Predictions by Model

We can see that in all cases, the BMA model was able to generate posterior distributions with the narrowest credible interval for each candidate, giving us a higher degree of belief than in any of the individual models. Now that we have support that our BMA model has the most desirable credible intervals, we can move forward with this model to compare the overall election results.

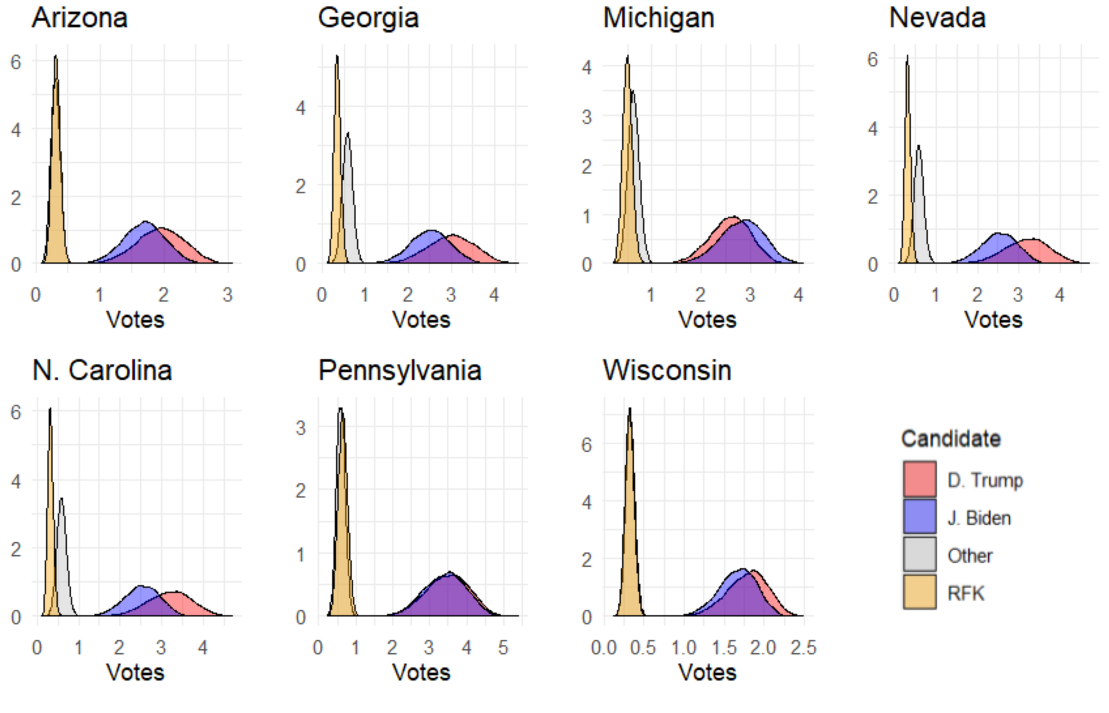


Figure 6: BMA Model Predictions for each State

Although the results between Trump and Biden appear very close in many states, it's important to remember that these curves follow a multinomial relationship, and are not independent from one another. This means that although the curves of Michigan, for example, show just a small sliver where Biden receives more votes than Trump, this means Biden is still receiving more Michigan votes than Trump in a great majority of simulations. To better understand this, we have to compare the election results in each of the 8000 iterations, and see which percentage of iterations each candidate has more votes than the other.

State	**Biden (%)**	**Trump (%)**
Arizona	1.075	98.925
Georgia	0.2375	99.7625
Michigan	91.025	8.975
North Carolina	0.0	100.0
Nevada	0.0	100.0
Pennsylvania	37.6	62.4
Wisconsin	14.875	85.125

Table 6: Statewise Win Percentages for Biden and Trump

The statewise win percentages across simulated iterations show that many of the states are more decisive than the density plots may make it seem. Many states have one candidate winning in nearly or exactly 100% of the iterations. The three most competitive states are Michigan with Biden winning the majority vote 91.025% of the time, Pennsylvania with Trump winning 62.4% of the time, and Wisconsin with Trump winning 85.125%.

These statewise win percentages are insightful, but the ultimate goal here is to predict the results of the national election considering the electoral college system described in section 1.3. To do this, we need to do a similar procedure as the one used to produce table 6, but now keeping track of the winner of each state in each iteration, and then giving the winner a set amount of electoral points according to which states they won. Because this model only includes seven of the most competitive states, we are going to assume that to begin, the rest of the states have already been contributed their points towards whichever candidate is polling ahead. This grants Biden 229 points to begin, Trump 216, and 0 points for any other third-party option. Summing all the electoral college points for each candidate gives us the following totals across 8000 iterations:

```
> election_wins
Biden Trump Other   RFK
  417  7582     0     0
```

Meaning that trump is winning more electoral votes, and therefore the presidency, in 7582 of the simulations, or 94.775% of the time. 5.2125% of the time, Biden wins.

If you're attentive you'll see that these sums do not add up to 8000. 7852 Trump wins and 417 Biden wins gives us with 7999 total iterations. So what happened to the 8000th iteration? Because there are a total of 538 electoral votes, it is technically possible for a tie to occur. We can check to see if there was a tie in any iteration by searching for how many times each candidate received 269 votes.

```
> colSums(electoral_college_results == 269)
Biden Trump Other   RFK
    1     1     0     0
```

This explains the missing 8000th iteration. So, in 1 out of 8000, or 0.0125% of the simulations, Trump and Biden reached an exact tie in the electoral college. Although this is rare, this has happened before in the United States, and there is a plan for this, as it has happened exactly once, in 1800. The US Constitution[6] specifies the following instructions in case of an electoral tie:

> *"If there be more than one who have such Majority, and have an equal Number of Votes, then the House of Representatives shall immediately chuse by Ballot one of them for President."*

Given that the House of Representatives currently has a Republican majority, it's most likely that this electoral tie would result in a Trump victory, although it can't be assumed for certain that all representatives would vote in allegiance along party lines.

A few more ways to visualize the final electoral point summation may be helpful, so the following table offers a basic statistical summary of the total electoral points scored by each candidate. It's worth noting that that the median and 1st quartile values are the exact same for Biden, and the median and 3rd quartile for Trump (because their scores have an inverse relationship). These quartiles and medians being the same indicates the results are somewhat concentrated around a central value, without extremely large variability.

Statistic	**Biden**	**Trump**	**Other**	**RFK**
Minimum	229.0	238.0	0	0
1st Quartile	244.0	275.0	0	0
Median	244.0	294.0	0	0
Mean	251.4	286.6	0	0
3rd Quartile	263.0	294.0	0	0
Maximum	300.0	309.0	0	0

Table 7: Summary of Electoral Votes for Each Candidate

[6] *Constitution of the United States*, Art. II, Sec. 1, 1787.

Finally, we can plot the density curve for the electoral points of each candidate, although its interpretation isn't very intuitive.

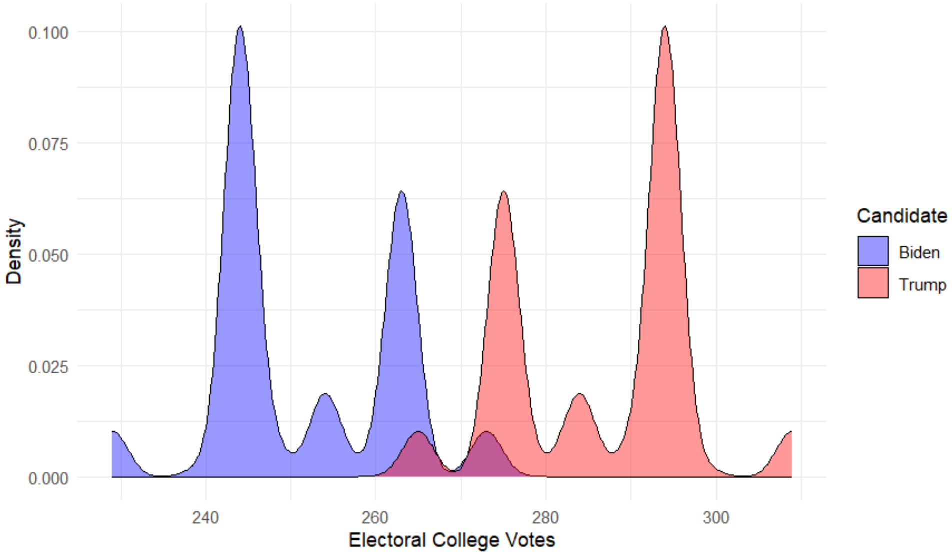


Figure 7: Density of Electoral Votes for Biden and Trump

Because the electoral points operate on a binary threshold basis, that is, they're either given to Trump or Biden based on the winner of the state, the density curve has a pretty unique shape. Each peak represents a different scenario in which a state was won or lost by a candidate, changing the total sum of electoral votes. And because the electoral votes of each candidate are not independent of the other, Biden and Trump's density distributions are actually just mirrors of one another. It can be seen that Trump is winning the vast majority of the simulations, and the two curves only intersect in a small purple region around 270 votes, which is the required votes to win. Only in the right half of this symmetric intersection is the region in which Biden receives more electoral votes than Trump, winning the presidency.

6 Discussion

While Bayesian Model Averaging is a very powerful tool, many of the potential hazards associated with Bayesian statistics are amplified with its implementation. In section 4 of this paper, we saw that different choices for model assumptions can completely change the outcome of BMA, and every decision must be thoroughly justified. However, even in the worst case scenario for my models, modern marginal likelihood computation methods can serve as robust selection tools to choose and eliminate models based on their likelihood. In cases of varying leniency to model assumptions, BMA has shown to compound agreeing posteriors into tighter credible intervals, and in the case of conflicting models, create meaningful compromises.

We conclude that stratification by gender is the best demographic to predict election turnout rates by, and by supplementing these predictions with age and race stratifications, we can create higher certainty predictions for election outcomes.

Although I've been using the term 'prediction' liberally throughout the paper, I should express caution about the interpretation of these 'predictions', and I think it's important to keep in mind the full question that was asked in the Morning Consult polls:

> *If the November 2024 election for U.S. president were held today, and Democrat Joe Biden, Republican Donald Trump, Independent Robert F. Kennedy Jr., Independent Cornel West, and Green Party candidate Jill Stein were on the ballot, for whom would you vote?*

That is, these polls and predictions represent the voting intentions and public opinion of voters *today.* For this reason, I don't propose the utility from models like this to be in the forecasting of future events, but rather tools to help us understand current sociology.

The electoral college system is inherently more complicated to understand than the majority of systems used in the world for the election of national presidents. Although right now, as of June 2nd 2024, many national election polls show Biden and Trump winning a very close amount of votes nationwide at around 44.1%, hinting at a close election, a more practical interpretation of election polls is needed to understand how these polls can transfer into electoral outcomes. With one candidate winning 94.8% of posterior samples, the current state of public opinion reflects a much more lopsided outcome than a quick glance at national polls would suggest.

This information is important and useful for a variety of reasons. Ordinary voters might be interested in the degree to which the elections in their state will impact the results of national elections and the competitiveness of said state. Newscasters and journalists hopefully feel an obligation to transmit honest representations of elections, requiring an analysis that goes deeper than simple voting percentages. The 2016 election became infamous for having incorrect polls that 'mislead' public opinion resulting in a nationwide shock when Trump was declared the winner. Although the national polls in 2016 actually weren't that far off, there were some serious issues with the state polls and their interpretation. Some issues included improper estimation of turnout rate, last-minute decision changes, and improper poll weights,[7] particularly weights relating to education level. Averaging different types of stratified weights through BMA offers a promising solution to reduce bias of these weights.

A final and equally important use for a practical Bayesian interpretation of polls is for candidates themselves. As clarified earlier, these 'predictions' made by my model are not so much probabilities of candidates winning at some point in the future as much insights of public opinion. With a more realistic understanding of their constituents, politicians can behave accordingly, whether it be to focus on the states or demographics needed to secure an election, or by using their existing political power to enact change that generates positive public support. With polls being published not just months, but years before elections take place, there is an infinite amount of unpredictable events that can occur between their release and elections. However, the usefulness they serve in understanding the present is immediate, it may just require a bit of data analysis.

[7] Steven Shepard, "5 takeaways from the 2016 polling autopsy," 2017, accessed June 2, 2024, https://www.politico.com/story/2017/05/04/2016-election-pollsters-react-237975.

References

Constitution of the United States. Art. II, Sec. 1, 1787.

Consult, Morning. “Swing States Tracking Poll #2404025: April 08-15, 2024.” Accessed June 1, 2024, 2024. Accessed June 1, 2024. https://pro-assets.morningconsult.com/wp-uploads/2024/04/Bloomberg_2024-Election-Tracking-Wave-7.pdf.

Forsberg, Ole J. *Understanding Elections through Statistics.* Boca Raton, Florida: CRC Press, 2020.

Gronau, Quentin F., Henrik Singmann, and Eric-Jan Wagenmakers. *bridgesampling: An R Package for Estimating Normalizing Constants.* Software manual. Accessed: 2024-06-01. The Comprehensive R Archive Network, 2017. Accessed June 1, 2024. https://cran.r-project.org/web/packages/bridgesampling/vignettes/bridgesampling_paper_extended.pdf.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. “Bayesian Model Averaging: A Tutorial.” *Statistical Science* 14, no. 4 (1999): 382–417.

Report, U.S. News & World. “7 Swing States That Could Decide the 2024 Presidential Election.” Accessed June 1, 2024, 2023. Accessed June 1, 2024. https://www.usnews.com/news/elections/articles/7-swing-states-that-could-decide-the-2024-presidential-election.

Shepard, Steven. “5 takeaways from the 2016 polling autopsy,” 2017. Accessed June 2, 2024. https://www.politico.com/story/2017/05/04/2016-election-pollsters-react-237975.