Forecasting the 2024 U.S. Presidential Election: Kamala Harris' Projected Victory*

A Bayesian Analysis of Polling Data from Key Battleground States"

Jimin Lee   Sarah Ding   Xiyan Chen

November 4, 2024

This paper presents a predictive model for the 2024 United States Presidential Election, focusing on essential battleground states. Utilizing a Bayesian framework and state-level polling data, we estimate the probabilities of victory for candidates Donald Trump and Kamala Harris. Our findings indicate that Harris is projected to secure a win in six out of seven key battleground states, such as Michigan, Nevada, and North Carolina, with predicted support exceeding 50%. This analysis not only forecasts the likely winner of the U.S. election but also equips society and the economy to better undertstand the current U.S. political climate.

Table of contents

---

1 Introduction

The 2024 United States Presidential Election is set to be a crucial moment in American political history, with battleground states serving as pivotal determinants of the electoral outcome. As candidates Donald Trump and Kamala Harris intensify their campaigns, understanding voter preferences in these key regions is essential for accurate election forecasting. This paper utilizes state-level polling data from FiveThirtyEight (FiveThirtyEight 2024) and the statistical programming language R (R Core Team 2023) to develop a predictive model estimating each candidate's likelihood of victory in critical battleground states, including Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin.

The primary focus of this study is the probability of victory for Trump or Harris in these battleground states, derived from aggregated polling averages. Employing a Bayesian hierarchical model combined with natural splines, our analysis effectively captures the temporal dynamics of voter sentiment, allowing us to estimate winning probabilities in these decisive areas. Our findings reveal a competitive landscape, with Harris projected to win in six of the seven battleground states analyzed, notably in Michigan and Nevada, where support leans in her favor. However, Trump remains competitive in Arizona, illustrating the variability of voter preferences across these regions.

These predictions are critical not only for forecasting the likely outcome of the election but also for enhancing societal and economic understanding of the current political climate in the U.S. The implications of our findings extend beyond mere predictions, offering a framework for analyzing potential shifts in voter sentiment as Election Day approaches. The remainder of this paper is organized as follows: Section Section 2 details the data and variables utilized in the analysis; Section Section 3 outlines the model setup and estimation strategies employed; Section Section 4 presents the results, including visualizations of the predicted election outcomes; and Section Section 5 discusses the implications and limitations of our findings.

The analysis, modeling, and visualizations of this paper were conducted using the programming language R (R Core Team 2023) and the following packages: `tidyverse` package of Wickham et al. (2023) , `janitor` package of Firke (2023) for data cleaning, `kableExtra` package of Zhu (2023) for dynamic report generation and table formatting, `modelsummary` package of Arel-Bundock (2023) for summarizing models, `usmap` package of Albers et al. (2023) for mapping U.S. states, `arrow` package of Richardson et al. (2023) for data integration with Apache Arrow,

`lubridate` package of Grolemund and Wickham (2011) for date handling, `rstanarm` package of Gabry et al. (2023) and `brms` package of Bürkner (2023) for Bayesian regression modeling, `Matrix` package of Bates and Maechler (2023) for handling sparse and dense matrices, `splines` package of Team (2023) for regression splines, `caret`package of Kuhn et al. (2023) for machine learning support, and `broom.mixed` package of Bolker et al. (2023) for tidying mixed models.

2 Data

2.1 Overview

The data for this paper contains the polling information of the 2024 Presidential Election cycle and was obtained from Project 538's polling project (FiveThirtyEight 2024) on October 19th, 2024. The data is periodically updated throughout the presidential election campaign to reflect current and most up-to-date polling results. It consists of polling information from various pollsters such as the New York Times, ActiVote, Morning Consult, etc. For every poll conducted, relevant variables such as **'state'**, **'start_date'**, **'end_date'**, **'transparency_score'**, **'candidate_name'**, and percentage of support (**'pct'**) for the appropriate candidates are included. The data compiles information from all major pollsters in the US and provides us with an understanding of the current state of the presidential election.

The present analysis focuses on Donald Trump and Kamala Harris, the two leading candidates in the current presidential election race. For all the variables mentioned (see Section 2.3), we collected observations for both of the candidates. We collect observations from the data where the **'candidate_name'** is Trump and Harris. We are interested in the support for both candidates over time, for each state, therefore, we collect observations of **'end_date'**, and **'state'**, as well as the electorate support for each candidate is denoted as **'pct'**. Furthermore, since there are many pollsters present in the data, we extracted pollsters with high-quality data that is reliable, which is denoted as having a high **'numeric_grade'** and **'transparency_score'**, details of data cleaning are further explained in Appendix A.

2.2 Measurement

Voters' opinions are multifaceted and can be influenced by numerous factors, including media coverage, recent political events, personal beliefs, and social pressures. Each voter's decision is shaped by a unique mix of these factors, which can change frequently, especially close to an election. Pollsters start by converting the complex, multi-layered opinions in voters' heads into a simple polling response. This is the first challenge: each voter's thoughts might not fit neatly into options like "Trump," "Harris," or "Undecided." People may lean toward one candidate but have reservations, or may change their views over time. By asking for a single answer, the poll essentially "forces" an opinion that may not fully capture the voter's internal conflict.

Once collected, these individual responses are combined into percentages of support. This aggregation abstracts further, treating all "Trump" responses equally, regardless of how confident or reluctant the voter might have been. In effect, the complexity of each opinion is smoothed over into a uniform data point, with many variations of opinions lost in the process. Finally, these support percentages are fed into a model to make a predictive statement about the election outcome. The model's goal is to determine who is likely to win, based on the assumption that the aggregated opinions in our data reflect real-world voter behavior on Election Day. However, this step relies heavily on simplification, assuming that the snapshot of opinions captured in the polls will hold steady and that all voters will follow through on their expressed preferences. The raw polling data gives us percentages of support for each candidate in each state. To measure our outcome variable (the likelihood of winning), we convert these percentages into a binary outcome in our model (predicting Trump or Harris as the winner based on who has higher support in each state). Our model uses these support percentages to generate a probability of each candidate winning. This transformation helps translate fluctuating polling data into a clear prediction of the likely winner, but it also involves a level of simplification, as it assumes these support percentages accurately reflect voting intentions.

Using state as a predictor variable, which influences our outcome variable, assumes that all responses within a state are comparable and that regional differences are consistent across the state. However, each state has internal diversity—urban vs. rural areas, differences in income levels, or cultural values—that the variable "state" may not fully capture. This simplification is necessary for the model, but it means that nuanced, within-state variations in opinion are lost in the abstraction. Another predictor variable, end_date is also measured, it is designed to capture how voter sentiment changes over time, especially as the election approaches. By using the end date of each poll, the model can recognize patterns or shifts in support due to campaign events, news coverage, or debates. While "end_date" provides temporal context, it simplifies complex, often continuous changes in opinion into a discrete snapshot. This abstraction assumes that voter sentiment is relatively stable between poll dates, which may not reflect reality. For example, an event occurring immediately after a poll's end date could dramatically shift opinions that the model wouldn't capture until the next poll. Additionally, the importance of time in influencing voter sentiment is assumed to be similar across states, but certain states might react differently to events depending on their local context or demographic makeup.

2.3 Variables

The collected data from Project538 (FiveThirtyEight 2024) contains several key variables relevant to the analysis:

- **'State'**: Indicates the state where the polling took place, critical for understanding regional support dynamics.
- **'Polling Date'**: The date when the poll was conducted, which helps in analyzing trends over time.

- **'Pollster'**: The organization conducting the poll, providing insight into the reliability and methodology of the polling data.
- **'N'umeric Grade'**: A score assigned to each poll based on its methodology and historical accuracy, which assists in filtering out less reliable polls.
- **'Transparency Score'**: A measure indicating how transparent the pollster is regarding their methodology, further enhancing data credibility.
- **'Candidate Name'**: Identifies the candidate being polled, specifically focusing on Donald Trump and Kamala Harris.
- **Percentage of Support (pct)**: Represents the percentage of respondents supporting each candidate, which is the primary outcome variable of interest.

To ensure the quality of the analysis, only those polls that meet specific thresholds were included - for more details on the data cleaning process, refer to Section A.1.

2.3.1 Outcome Variable

As our focus is to analyze and predict voter support for Kamala Harris and Donald Trump during the current presidential election cycle, the primary outcome variables of interest are the percentages of support for each candidate, denoted as 'pct'.The pct is categorized by candidate and scaled to a proportion by diving by 100. This variable represents the proportion of voter support for each candidate based on polling data. Through understanding the percentage of support (pct) for each candidate, we can observe and infer the candidate that has a leading advantage in winning the presidential election race.

To give an overview of the outcome variable, divided per candidate, we provide summary statistics of this key variable, highlighting the characteristics of the polling support for each candidate across the battleground states.

Table 1: Summary Statistics for Trump and Harris Polling Data - Presents key summary statistics, including total polls, average support percentages, and the maximum and minimum polling levels for both candidates across the specified battleground states. All values are rounded to three decimal places, providing an overview of the polling landscape as of the selected election date.

Statistic	Value
Total Polls	107.000
Average Trump Support	47.801
Max Trump Support	51.700
Min Trump Support	43.000
Average Harris Support	47.999
Max Harris Support	51.200
Min Harris Support	43.600

Table 1 reveals a total of 107 polls conducted, calculated as the total count of rows in our cleaned dataset. 'Total polls' reflects the total number of unique poll entries after filtering for only battleground states, polls with sufficient data quality, and polls that report support percentages for both Trump and Harris. The 107 polls conducted display an average support level of 47.801% for Trump and 47.999% for Harris. This suggests that based on the dataset, Trump has a leading advantage in winning the presidential election race by having a slightly higher average level of support than Harris. The highest percentage of support for Trump in the 107 polls was 51.700% while the lowest recorded support was 43%. For Harris, the highest percentage of support also reached 51.200%, with a minimum of 43.600%. These statistics illustrate the competitive landscape of the presidential election race, without a clear leading candidate in the current race.

It is important to note that since **'pct'** stands for percentage, there may be a misconception that the values for Trump and Harris will always add up to 100%. For instance, in the summary statistics presented in Table 1, the average support for Donald Trump is 47.801%, while the average support for Kamala Harris is 47.999%. This results in a combined average support of 95.800%, indicating that there is a significant portion of respondents who are either undecided or supporting other candidates. This highlights the competitive nature of the election and emphasizes the necessity of examining not only the percentages attributed to the main candidates but also the context of voter preferences that could influence the final outcome.

2.3.2 Predictive Variables

In this section, we outline the two key predictor variables used in our analysis: **'state'**, which captures the geographical context of the polling data, and '**end_date**', which reflects the timing of the polls relative to the election, allowing us to assess trends in voter support over time.

2.3.2.1 State

The state (**'state'**) variable indicates the U.S. state where the poll was conducted or targeted. In the context of the U.S. presidential election, the outcome is frequently determined by several key battlegroud states. In the current 2024 election cycle, there are seven battleground states that are pivotal in the contest between Kamala Harris, representing the Democratic Party, and Donald Trump, representing the Republican Party. To enhance our model and refine our predictions, we focused exclusively on these seven battleground states. The polling results from each of these states play a crucial role in predicting electoral outcomes, as the percentage of support is evaluated on a state-by-state basis. The distribution of the state variable is visualized in Figure 1 and indicates that some states, such as Pennsylvania, North Carolina and Wisconsin have more polling observations compared to others, like Nevada. This variability reflects differences in polling frequency and interest among these battleground states. States
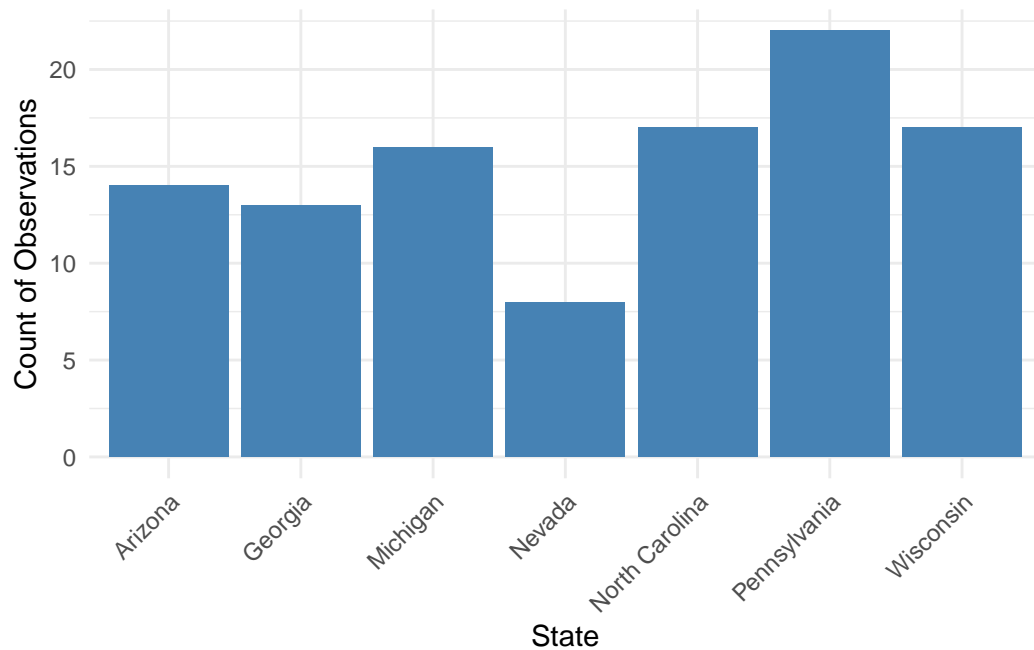
Figure 1: Distribution of Observations for Each State in Polling Data. Illustrates the number of polling observations across key battleground statesn. Pennsylvania has the highest count of observations, suggesting a more robust polling presence, while Nevada shows fewer observations.

with more polling observations provide a larger sample size, potentially leading to more stable and reliable predictions for those states in the model.

2.3.2.2 End Date

End date (**'end_date'**) represents the date the poll ended, essentially the date the poll stopped collecting data from electorates. In the context of our paper, the end date of each poll serves as one of the predictive variables and is used to assess changes in voter support over time. Including end date as a variable in our model can help us identify patterns in voter support as Election Day approaches, potentially reflecting the impact of campaign events, debates, or other political developments. Figure 2 displays the distribution of polling end dates from July 2024 to October 2024.
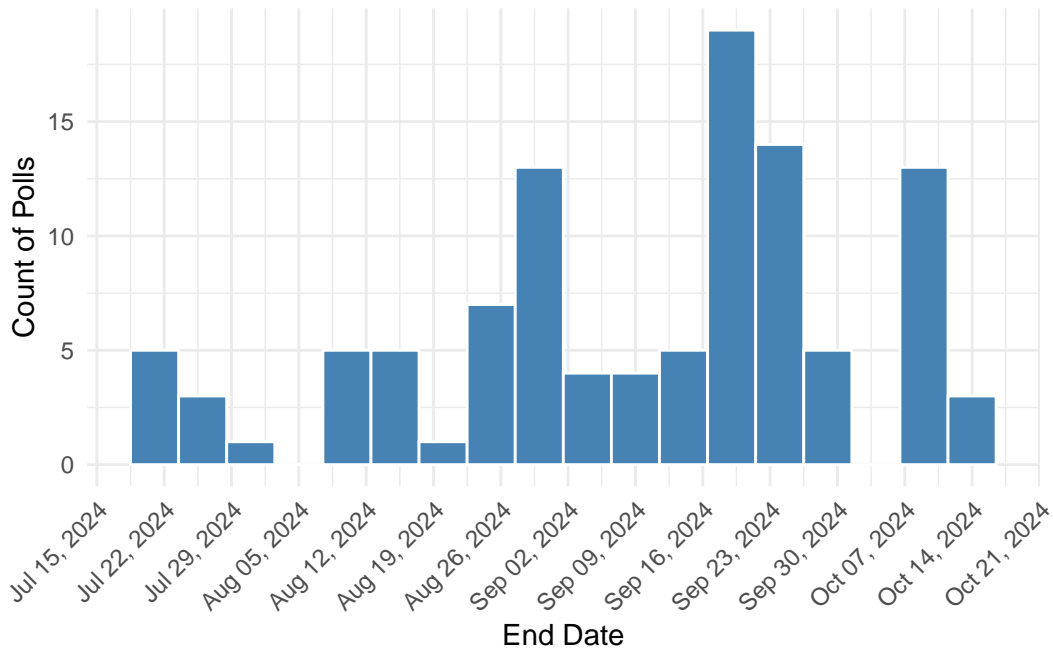
Figure 2: Distribution of Polling End Dates. This histogram displays the frequency of polling end dates for key battleground states, highlighting peaks in polling activity around late September.

We filtered for the earliest end date as July 2024 since Harris officially announced her presidential campaign on July 21, 2024. Therefore, any polls that has an end date before July would not contain polling results for Harris. The data from (FiveThirtyEight 2024) is last updated on October 15, 2024, thus the latest end date being October 21, 2024. Furthermore, Figure 2 also reveals that polling activity is not evenly distributed over time. There are clear peaks where the number of polls conducted is higher, notably in early September, late September, as

well as early October. This concentration of polls likely corresponds to period of heightened political interest or significant campaign events that prompt increased polling.

2.3.3 Other Variables

This section details additional variables that provide essential context for our analysis, including **'numeric_grade'**, which evaluates the quality of the polls; **'transparency_score'**, indicating the openness of poll methodologies; **'pollster'**, identifying the organizations conducting the polls; and **'start_date'**, which denotes the initiation of the polling period for each survey.

2.3.3.1 Numeric Grade

Numeric Grade (**'numeric_grade'**) represents the quality of each poll, as rated by (FiveThirtyEight 2024), based on factors such as methodology, transparency, and historical accuracy. In the context of our paper, numeric grade allows us to weigh the reliability of different polls. By including numeric grade as a variable in our model, we aim to account for variations in poll quality, ensuring that polls with higher ratings (indicating higher reliability) contribute more significantly to our predictions of voter support.

Figure 3 displays the distribution of numeric grades for the polls included in our analysis, ranging from 2.5 to 3.0. Polls with higher grades are generally seen as more methodologically rigorous and transparent, which helps improve the reliability of our model's predictions. The distribution shows several peaks, indicating that certain grade levels (around 2.7, 2.8, and 3.0) are more common among the polling data we collected. This variability in numeric grade highlights the diversity in polling quality within our dataset. The presence of peaks suggests that some pollsters consistently receive higher ratings, likely due to their use of more rigorous methods or transparent reporting standards. Conversely, the valleys in the distribution may indicate fewer polls from sources with lower numeric grades, reflecting a potential tendency for lower-quality polls to be less prevalent or less frequently conducted. By accounting for numeric grade, we can improve the accuracy of predictions by emphasizing data from more reliable sources. This consideration is crucial, as polling quality can impact the stability of the model's forecasts, especially in closely contested states where polling accuracy is essential for reliable predictions.

2.3.3.2 Transparency Score

Transparency Score (**'transparency_score'**) represents the degree of methodological transparency provided by each pollster, rated on a scale from 0 to 10 by (FiveThirtyEight 2024). This score reflects the extent to which pollsters disclose details about their data collection and analysis methods, such as sample size, weighting, and margin of error. In the context of our paper, transparency score is an essential variable for evaluating the reliability of polling data,
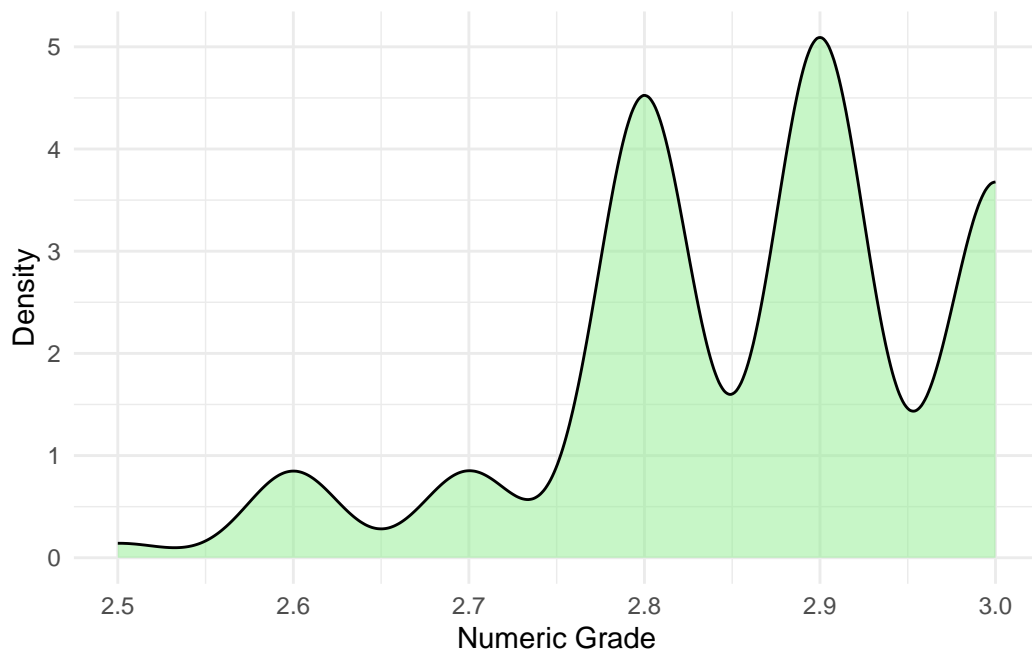
Figure 3: Density plot illustrating the distribution of numeric grades assigned to pollsters within the polling data. Highlights the frequency of numeric grades, indicating key concentrations that suggest variability in pollster quality.

as higher transparency scores generally indicate a greater level of methodological rigor and openness. Figure 4 displays the distribution of transparency scores across the polling data for key battleground states. The plot shows that transparency scores cluster heavily around ratings of 7 and 9, suggesting that a majority of polls in our dataset come from sources that maintain a relatively high level of transparency. There are smaller peaks at lower scores (5 and 10), indicating that while some pollsters offer limited transparency, a few polls also achieve perfect transparency scores.
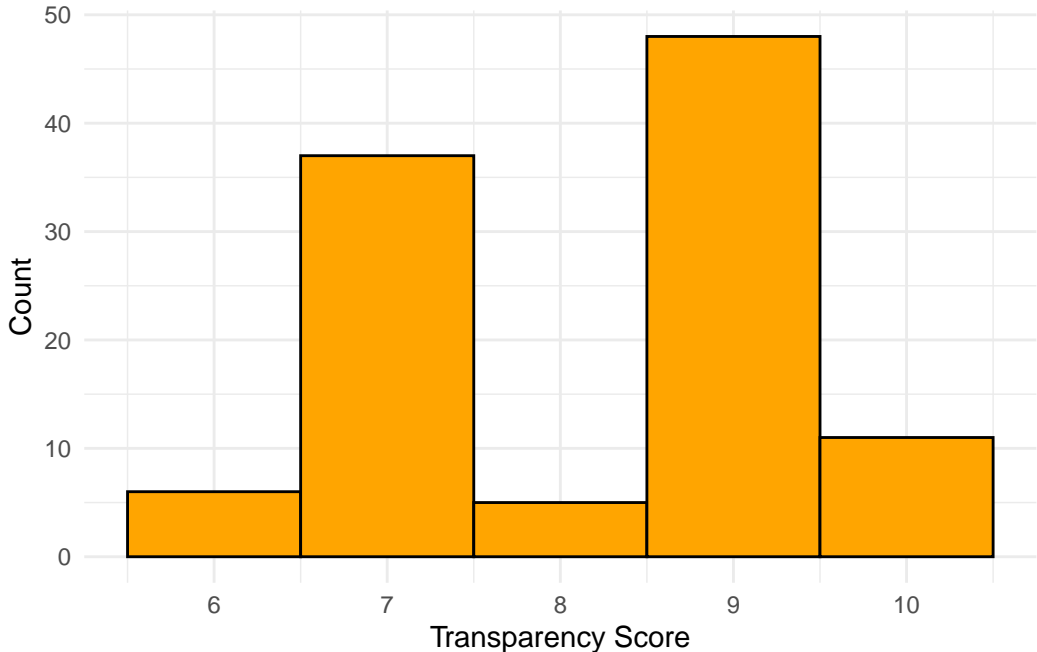
Figure 4: Histogram displaying the distribution of transparency scores among pollsters in the polling data. Indicates the frequency of various transparency scores, highlighting a concentration of scores around 7 and 9, which suggests a predominance of high transparency among the assessed pollsters.

This distribution is important for understanding the credibility of our data sources. Polls with higher transparency scores are more likely to provide consistent and reliable results, as they openly disclose their methodologies, allowing us to assess potential biases or limitations in their data. In contrast, polls with lower transparency scores may lack sufficient methodological detail, making it challenging to gauge their accuracy. Incorporating transparency scores into our model helps differentiate between high- and low-transparency polls. This adjustment is crucial as transparency can influence the stability of the model's predictions, especially in closely contested states where the reliability of polling data may impact the accuracy of our forecasted election outcomes.

2.3.3.3 Pollster

Pollster (**'pollster'**) represents the various organizations responsible for conducting polls across key battleground states in the lead-up to the 2024 United States Presidential Election. In the context of our paper, the diversity and frequency of polling by different pollsters are essential for understanding the reliability and variance in the data. Each pollster may use different methodologies, sample populations, and data collection techniques, which can affect the results and introduce variability in polling outcomes.
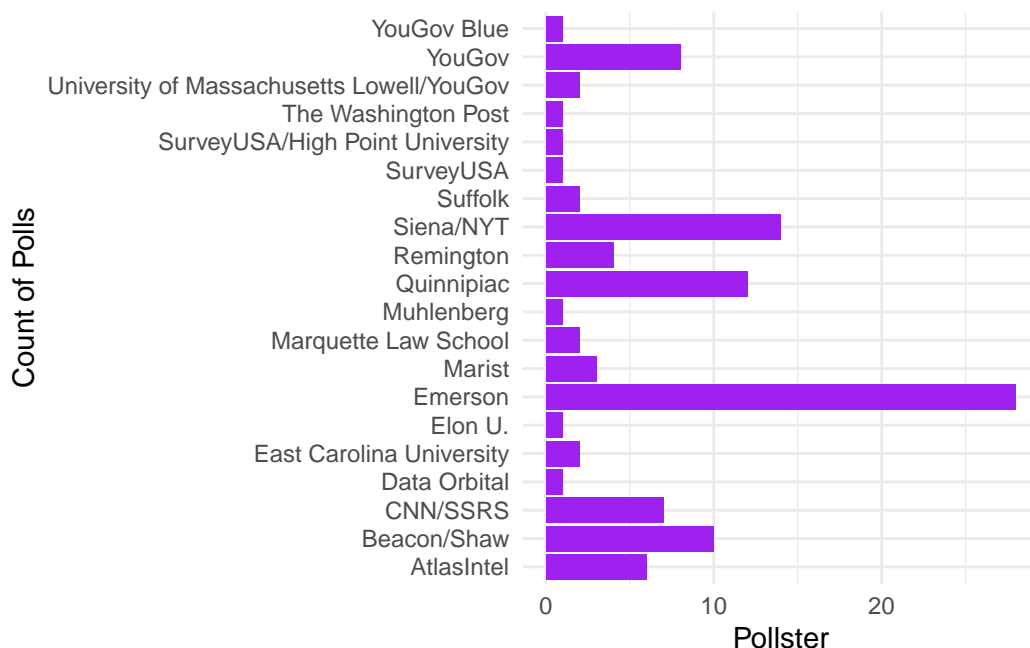
Figure 5: Horizontal bar chart displaying the distribution of polls conducted by various pollsters in key battleground states. Highlights the frequency of polling activities across different organizations, with Marist and Emerson standing out as the most prolific pollsters.

Figure 5 displays the count of polls conducted by each pollster, illustrating the distribution of polling efforts across multiple organizations. Notably, Emerson is the most frequent pollster, conducting the highest number of polls, followed by Quinnipiac and Siena/NYT. Other pollsters, such as YouGov, CNN/SSRS, and Marist, also contribute a significant number of polls, whereas some organizations like The Washington Post and SurveyUSA have comparatively fewer entries in the dataset.This distribution is crucial for evaluating the consistency and breadth of the polling data. Pollsters with a higher count of polls, like Emerson, contribute substantially to the dataset and can have a greater influence on the model's predictions, particularly if their methodology or sample selection differs from others. On the other hand, pollsters with fewer entries may provide valuable but limited data points, potentially adding

unique perspectives but with less influence on overall trends.

2.3.3.4 Start Date

Start date (**'start_date'**) represents the date when each poll began collecting responses. In the context of our paper, start date is a key variable as it allows us to focus on polling data collected after July 21, 2024—the date Kamala Harris officially announced her presidential campaign. By filtering the dataset to include only polls with a start date after this announcement, we ensure that each poll reflects voter sentiment with Harris as an active candidate, enabling a more accurate comparison between her and Donald Trump. Figure 6 displays the distribution of polling start dates from late July to mid-October 2024, showing when polling efforts began across key battleground states. The histogram reveals peaks in polling activity, particularly in mid-August, early September, and late September, indicating periods of heightened interest or campaign events that likely prompted increased polling.
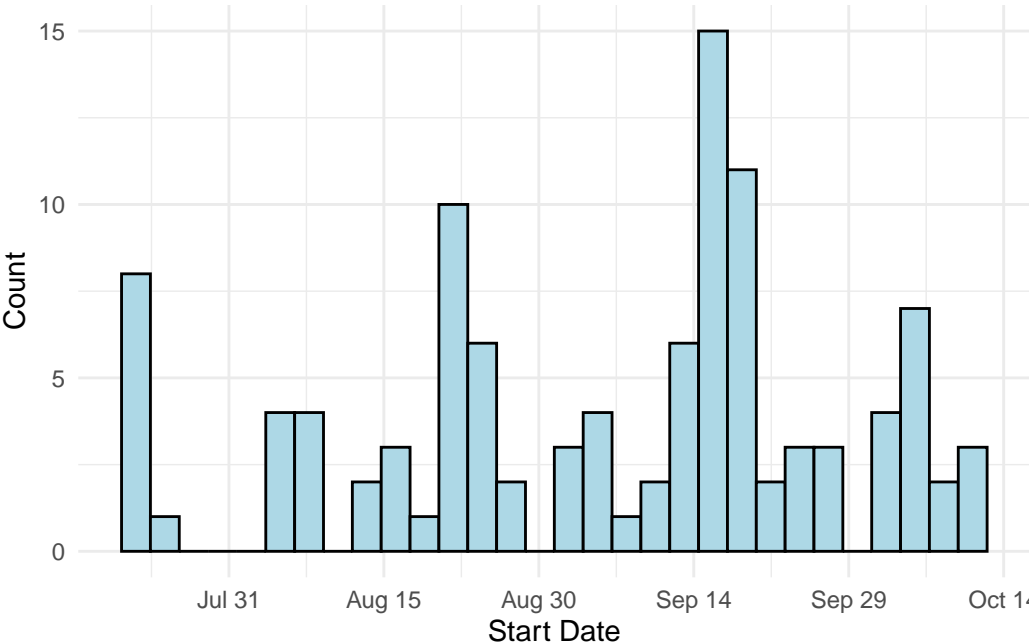
Figure 6: Histogram illustrating the distribution of polling start dates for key battleground states. Displays the frequency of polling activities initiated over time, with a notable peak on September 14, indicating a surge in polling efforts as the election approached.

This distribution provides insights into the timing and intensity of polling efforts throughout the campaign season. By including only polls initiated after Harris's campaign launch, we capture voter preferences in a context where both major candidates are actively campaigning, thus making the polling data more relevant for direct comparisons. This filtered timeline

ensures our model focuses on data that reflects the competitive dynamics between Harris and Trump, enhancing the validity of our predictive analysis as we approach Election Day.

3 Model

3.1 Model Overview

The Bayesian Beta regression models constructed in this paper aim to predict the probability of victory for Donald Trump and Kamala Harris in the 2024 U.S. Presidential Election by analyzing polling data of the 7 battleground states. One Bayesian Beta regression model is constructed for each candidate, and each of them estimates the proportion of support for each candidate based on polling data, accounting for both polling trends over time, and state-level differences in voter sentiment. These models allow us to evaluate how support for each candidate fluctuates over time in key states, and predict potential electoral outcomes.

One of the goals of our model is to analyze changes in voter support for each candidate over time, as Election Day approaches. By using end_date_num (days the earliest poll date) as a predictor with natural splines, the model captures non-linear trends in support. The time-based analysis provides us with information about how voter preference may shift due to campaign events, news cycles, and other influential factors in the months leading up to the election.

Another goal of our model is to predict the potential winner of the Presidential Election. We recognize that battleground states have unique political landscapes, so we include state as a random effect in each model. This allows each of the 7 battleground states to have a unique baseline level of support for each candidate, capturing regional differences in voter preferences. By incorporating state-level effects, the model can show variations in support across different states, revealing the potential winning candidate for the presidential election. Further background details, model specifications, and diagnostics are included in Appendix B.

3.2 Model Assumptions

In order to ensure the validity and reliability of our Bayesian models for predicting voter support percentages for Donald Trump and Kamala Harris, we establish several key assumptions that guide the modeling process:

- **Linearity in the Log-Odds**: The relationship between predictor variables (end_date and state) and the log-odds of the outcome variable (percentage of support) is linear. To account for non-linear trends over time, nature splines are used.
- **Independence**: The outcome for each polling observation is assumed to be independent of others. Each poll represents a distinct sample of voter sentiment at a specific time and location.

- **Appropriate Distribution for Proportions**: Since the model predicts support percentages (bounded between 0 and 100), a Beta distribution with a logit link is applied, which is ideal for data that are restricted to this range.
- **Random Effects**: To capture variation across states, state-level random effects are included, allowing each state to have its unique baseline level of support, acknowledging regional political differences
- **Prior Distributions**: The model uses weakly informative priors for the intercept and coefficients, which helps in generating stable estimates and reducing the influence of outliers or unusual polling data.

3.3 Model Setup

Our models were ran in R (R Core Team 2023) using the 'rstanarm' package of Gabry et al. (2023).

3.3.1 Trump Model

The primary estimand for Donald Trump's model is the probability of victory in the selected battleground states, derived from a logistic regression framework. The model is structured as follows:

$$P(\text{Victory}_{Trump} = 1) = \frac{e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}{1 + e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}$$

Where:

- $P(\text{Victory}_{Trump})$ is the predicted probability of Donald Trump winning.
- β_0 represents the intercept, indicating the baseline log-odds of winning.
- β_1 is the coefficient for the natural spline transformation of the numeric end date (end_date_num), capturing nonlinear trends in voter support over time.
- u_{state} accounts for random effects specific to each state, reflecting variations in voter preferences.

The dataset for the Trump model was prepared by extracting the relevant polling data, including the state, end date, and corresponding percentage of support for Trump ($Trump\_pct$). The end date was converted into a numeric format, denoting the number of days since the earliest polling date in the dataset. This transformation allows the model to effectively capture the evolving nature of voter sentiment as the election approaches.

The model was implemented using the 'brms' package (Bürkner 2023) in R (R Core Team 2023), applying a Beta distribution with a logit link function to accurately model the proportions of support:

$$y_i \sim \text{Beta}(\mu_i, \phi)$$

Here, y_i is the proportion of support for Trump in a given state, μ_i is the mean of the Beta distribution, and ϕ is the precision parameter reflecting variability.

We specified priors for the model parameters as follows:

- For the intercept β_0, we set a prior of $\beta_0 \sim \text{Normal}(0, 10)$, allowing the data to inform the estimation of baseline log-odds without imposing strong assumptions.
- For the slope β_1, we applied a prior of $\beta_1 \sim \text{Normal}(0, 5)$, reflecting reasonable expectations regarding the influence of polling trends.

Natural splines were used to model the effect of the end date, accommodating observed non-linearity in voter support. Key assumptions include the independence of observations within states and the suitability of the Beta distribution for modeling proportions. However, limitations may arise from the accuracy of polling data and the assumption that historical voting patterns will predict future behavior.

3.3.2 Harris Model

Similar to the Trump model, the estimand for Kamala Harris's model is the probability of her victory in the selected battleground states, modeled within a logistic regression framework. The formulation is analogous:

$$P(\text{Victory}_{Harris} = 1) = \frac{e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}{1 + e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num} + u_{\text{state}}}}$$

Where the components mirror those described for the Trump model, tailored to reflect the probability of victory for Harris.

For this model, the dataset was constructed similarly, extracting polling data specifically for Harris ($Harris\_pct$). The same transformations were applied, converting end dates into a numeric format to facilitate the modeling of trends over time.

The Harris model was also implemented using the **brms** package, employing a **Beta distribution** with a **logit link function** to represent the proportions of support accurately. The same mathematical representation and priors were applied as in the Trump model:

$$y_i \sim \text{Beta}(\mu_i, \phi)$$

With the same definitions for y_i, μ_i, and ϕ.

Natural splines were employed to adapt to the non-linear trends in voter support over time. The assumptions and limitations discussed for the Trump model apply equally to the Harris model.

By utilizing Bayesian beta regression models for both candidates, we leverage their flexibility and capacity to incorporate prior beliefs, which is particularly advantageous in the context of electoral forecasting where uncertainty plays a critical role.

3.4 Model Justification

The Bayesian Beta Regression model with a logit link function is the most suitable model for predicting voter support for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election. Our primary outcome variable is the percentage of support for each candidate in seven key battleground states, which is naturally bounded between 0% and 100%. The Beta distribution is ideal for modeling this proportion data, as it allows us to avoid unrealistic predictions outside the (0, 1) range. To capture non-linear trends in voter support as Election Day approaches, we employ natural splines with **end_date_num**, which represents the number of days since the earliest poll date. Additionally, incorporating state-specific random effects acknowledges the variations in voter preferences across different regions, allowing for more accurate modeling of baseline support levels. The Bayesian framework enhances our model by enabling the incorporation of prior information, leading to more stable estimates, especially in the context of sparse polling data. Furthermore, the model's structure allows for clear interpretation in terms of victory probabilities, aligning well with our goals of forecasting electoral outcomes. Further details are discussed in Appendix B.

3.5 Model Summary

3.5.1 Trump Model Summary

Table 2 displays fixed effects estimates for several key terms that capture the influence of time on his support in the battleground states. The intercept, with an estimate of -0.0923, represents the baseline log-odds of Trump's support when other predictors are held constant. This negative intercept suggests a slightly lower baseline level of support. The term nsenddatenumdfEQ42, with an estimate of 0.0803 and a statistically significant positive confidence interval (0.0179 to 0.1426), indicates an increase in support around this time period. Similarly, nsenddatenumdfEQ44 has an estimate of 0.0789 and a confidence interval above zero (0.0014 to 0.1589), also showing a positive effect on support. Other terms such as nsenddatenumdfEQ41 and nsenddatenumdfEQ43 have less significant impacts, suggesting that certain time periods were more influential than others in shaping Trump's support. The standard errors for the fixed effects terms range from 0.0306 to 0.0676 for nsenddatenumdfEQ43. The relatively low standard error of 0.0306 on the intercept suggests that the baseline estimate is quite stable, meaning we can be more confident about this baseline support level. However, the standard

error for nsenddatenumdfEQ43 (0.0676) is higher than for other terms, indicating that this estimate is less precise.

Table 2: Coefficient estimates from the Bayesian Beta regression model for Donald Trump, showing the impact of the natural spline transformation of end date on voter support. The intercept indicates baseline log-odds, while spline terms reflect nonlinear trends in voter sentiment, with positive coefficients suggesting increased support as Election Day approaches.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.0922852	0.0305620	-0.1520050	-0.0333538
nsenddatenumdfEQ41	-0.0071493	0.0355882	-0.0793200	0.0623891
nsenddatenumdfEQ42	0.0802768	0.0317675	0.0179060	0.1426297
nsenddatenumdfEQ43	0.0025949	0.0676124	-0.1341383	0.1336160
nsenddatenumdfEQ44	0.0789177	0.0400326	0.0013945	0.1589037

These results imply that support for Trump varied over time in ways that may correspond to specific campaign events or other influences affecting voter sentiment. The significance of terms like nsenddatenumdfEQ42 and nsenddatenumdfEQ44 suggests that changes in voter preference could be captured effectively by this model. And by checking standard errors, helping us determine how much we can rely on those data.

3.5.2 Harris Model Summary

Table 3 similarly shows the impact of time on her support in the battleground states. The intercept for the Harris model is estimated at -0.1508, suggesting a slightly lower baseline support level compared to Trump(-0.0923). The term nsenddatenumdfEQ41 has an estimate of 0.1225 and a confidence interval 0.0553 to 0.1903, indicating a significant increase in support during this period. Another positive and significant effect is seen in nsenddatenumdfEQ43, with an estimate of 0.1191 and confidence interval from -0.0071 to 0.2417, which suggests a trend toward increased support. While other terms stay positive, they have less impacts on support. Additionally, the standard errors for fixed effects are generally smaller, ranging from 0.0279 for the intercept to 0.0631 for nsenddatenumdfEQ43. The larger standard error for nsenddatenumdfEQ43 (0.0631) suggests that there would be some uncertainty in how this time point influences support, though the positive estimate still points toward an upward trend. In comparison to the Trump model, the standard errors in the Harris model are slightly smaller for most terms, indicating potentially more consistent support trends over time in the data analyzed.

Table 3: Coefficient estimates from the Bayesian Beta regression model for Kamala Harris, illustrating the effects of the natural spline transformation of end date on voter support. The intercept represents the baseline log-odds, while the spline terms indicate nonlinear trends in voter sentiment, with positive coefficients suggesting an increase in support as Election Day nears.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.1508299	0.0279264	-0.2041163	-0.0954951
nsenddatenumdfEQ41	0.1225217	0.0341326	0.0552911	0.1903117
nsenddatenumdfEQ42	0.0248781	0.0291738	-0.0321764	0.0833965
nsenddatenumdfEQ43	0.1191128	0.0630871	-0.0070147	0.2416542
nsenddatenumdfEQ44	0.0669873	0.0382015	-0.0071597	0.1426828

These results highlight that Harris's support also fluctuated over time, with nsenddatenumdfEQ41 and nsenddatenumdfEQ43 marking periods of notable increases. Helping us understand the pattern, this model predicts Harris's probability of victory and provides insights for future changes in support as the election nears.

4 Results

In this section, we present the results of our analysis, focusing on the electoral support for Donald Trump and Kamala Harris across key battleground states during the 2024 Presidential Election cycle. We begin by examining the state-level polling averages for both candidates, which provide a snapshot of voter preferences as captured by various pollsters. Next, we explore how polling averages have evolved over time, highlighting trends that may influence election outcomes. We then assess the predicted probabilities of Trump winning on Election Day, followed by a detailed look at the likelihood of winning by state. Finally, we visualize the predicted winner by state on a US map, summarizing our findings and their implications for the electoral landscape.

4.1 State-Level Polling Averages for Trump and Harris

Table 1 displays overview of the percentage of support for Trump and Harris by each battleground state. It provides us with a high level understanding of the current voter dynamic. The pct results are aggregated averages of all pct data points for Trump and Harris in each state. By comparing the percentage of support for T.rump and Harris, we can determine the candidate that has a leading advantage in winning each state. For example,Trump has 49.2% support in Arizona while Harris has 46.8% support, which means Trump has won the state of Arizona.

Table 4: Percentage of support for Donald Trump and Kamala Harris across key battleground states. Trump maintains a slight lead in states such as Arizona and Georgia, while Harris shows stronger support in 5 of 7 key battleground states.

State	Trump %	Harris %
Arizona	49.2	46.8
Georgia	48.9	47.2
Michigan	46.9	48.1
Nevada	47.9	48.5
North Carolina	47.6	48.1
Pennsylvania	47.4	48.3
Wisconsin	47.2	48.9

This distribution is crucial for evaluating the consistency and breadth of the polling data. Pollsters with a higher count of polls, like Emerson, contribute substantially to the dataset and can have a greater influence on the model's predictions, particularly if their methodology or sample selection differs from others. On the other hand, pollsters with fewer entries may provide valuable but limited data points, potentially adding unique perspectives but with less influence on overall trends.

4.2 Polling Averages for Trump and Harris Over Time in Battleground States

Figure 7 provides a visual comparison of the current polling support for Donald Trump and Kamala Harris across key battleground states, tracked over time from August 2024 to October 2024. Each line plot represents the average polling percentages for both candidates in a specific state, with red lines denoting Trump's support and blue lines representing Harris's support. Across several states, there are noticeably increased fluctuations in polling averages for both candidates closer to the election particularly from late September to October. This pattern may indicate heightened voter interest and engagement as the election date approaches which is November 4th, 2024, it could also reflect changes in public opinion due to campaign events, political debates, or other external factors.

Out of the seven battleground states, Arizona and Georgia display a clear leading win by Trump while Nevada and North Carolina display a clear leading win by Harris. Michigan, Pennsylvania, and Wisconsin are the three states that don't have a clear leading winner. Therefore, these three states are the key states that could determine the election outcome as it approaches election day. Election outcomes could be easily flipped if either of the candidates can sweep more than one out of the three states. Candidates should focus on strategizing their campaigns to increase their voters' support at critical moments. The frequent fluctuations of greater margins also indicate the uncertainty and instability of voters' attitudes towards both candidates.
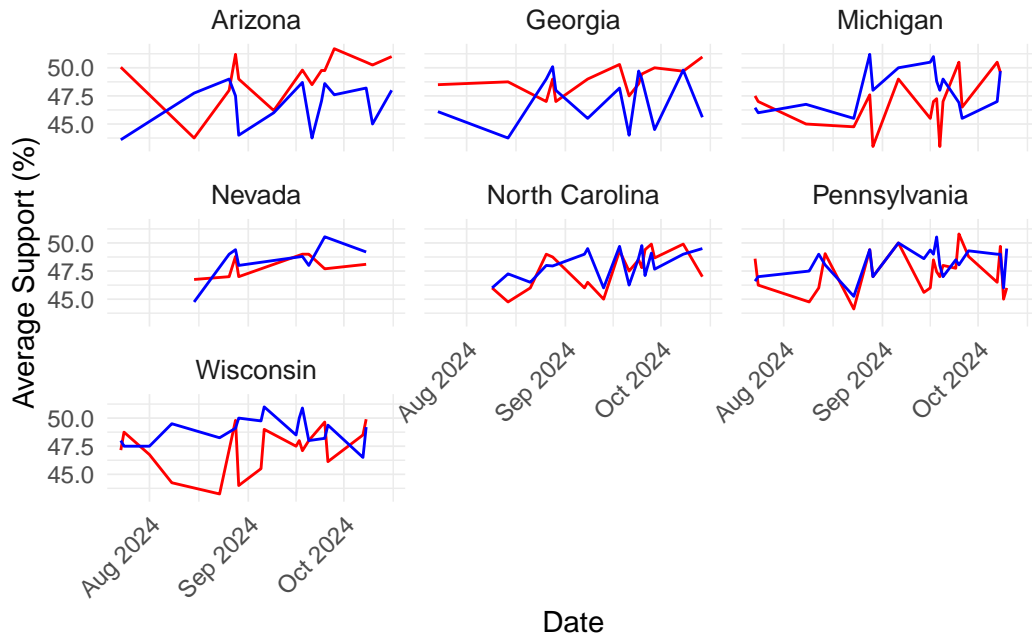
Figure 7: Polling averages for Donald Trump (red) and Kamala Harris (blue) in battleground states from August to October 2024. Captures the varying fluctuations in average support percentages across Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, indicating the competitive dynamics of the election.

Since it is difficult to conclude a definitive predictive outcome by observing the current data, we utilize our models, described in Section 3, to predict the possible outcomes by incorporating the given data. By using the current data, the model can identify the state-specific trends thus generating more accurate predictions of the outcome. Relevant prediction results are provided below in Figure 8.

4.3 State-Level Probability of Trump Winning on Election Day

By incorporating our Bayesian models, we generate the predicted probabilities of winner for both Donald Trump and Kamala Harris in key battleground states, on Election Day (**November 5th, 2024**). Figure 8 provides a visual summary of these predicted probabilities.
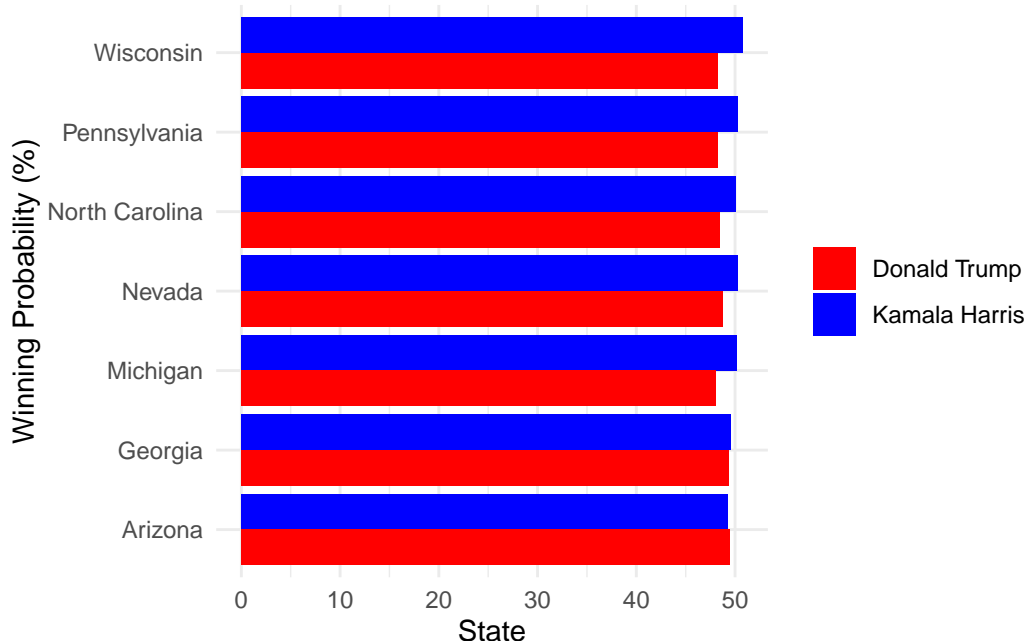
Figure 8: Predicted winning percentages for Donald Trump (red) and Kamala Harris (blue) in key battleground states on Election Day. Illustrates the estimated probability of victory for each candidate across Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, highlighting Harris's projected advantages in the majority of these states.

These predictions are based on polling data collected up to October 19th, 2024 to forecast each candidate's support within these critical states. Each state's political dynamics are unique, and this is reflected in the model's predictions. States like Arizona and Georgia show relatively close probabilities for both candidates, marking them as key states to focus on as they are

the most influential on the election outcomes. The five other states all show relatively distinct leading support for Harris, suggesting that Harris's support base in those states has been consistently strong in the polls leading up to Election Day.

4.4 Probability of Winning by State on Election Day

Table 5 displays the numeric percentage of the probabilities of winning for Trump and Harris. By comparing the predicted probabilities, the predicted winner for each state is assigned accordingly. For instance, the table displays that Trump has a 49.48% predicted probability of winning Arizona compared to 49.24% of Harris's, making Trump the predicted winner in Arizona. For all six other states, Harris has a higher predicted probability of winning than Trump.

Table 5: Predicted winning probabilities for Donald Trump and Kamala Harris in key battleground states, with percentages representing the likelihood of each candidate's victory. Summarized predicted support for each candidate in Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, indicating a competitive race, particularly with Harris expected to win in most states by narrow margins.

State	Predicted Trump Win (%)	Predicted Harris Win (%)	Predicted Winner
Arizona	49.48	49.23	Donald Trump
Georgia	49.34	49.54	Kamala Harris
Michigan	48.01	50.22	Kamala Harris
Nevada	48.79	50.29	Kamala Harris
North Carolina	48.49	50.06	Kamala Harris
Pennsylvania	48.27	50.31	Kamala Harris
Wisconsin	48.30	50.78	Kamala Harris

4.5 US Map Showing Predicted Winner by State on Election Day

To visualize the colour distribution of the predicted results, we emloy the US map from (Albers et al. 2023) to display the results from Table 5.

- 'Predicted Winner' column is Donald Trump: State filled red
- 'Predicted Winner' column is Kamala Harris: State filled blue

According to Figure 9, Arizona is the only state where Trump has a leading win while all the other six states are led by Harris. Since Harris has a six-to-one lead in the predicted percentage of support for all battleground states, we conclude that Harris is the predicted winner for the 2024 cycle of Presidential Election.

Figure 9: Predicted winners by battleground state on Election Day, illustrating the anticipated electoral outcomes for Donald Trump and Kamala Harris across key states. Highlights states where Trump is expected to win in red and states favoring Harris in blue, with gray indicating states where data is not applicable (NA). Underscores the competitive nature of the election and the distribution of predicted support across the country.

5 Discussion

5.1 Key Findings

This paper presents a forecast of the 2024 U.S, Presidential Election, leveraging Bayesian Beta regression models trained on polling data. We were able to model both the electoral support trend over time and the different political dynamics across the battleground states. The model's predictions reveal that, as Election Day approaches, Kamala Harris holds a lead in six of the seven battleground states–Pennsylvania, Michigan, Wisconsin, Nevada, North Carolina, and Georgia. However, her margins are consistently narrow, with predicted probabilities only slightly above 50% in each of these states. Harris shows minimal leads over Trump, with differences between the probability of winning of no more than 3 percent across all six states. Meanwhile, Trump also holds a narrow advantage in Arizona, suggesting a highly competitive electoral dynamic in this election.

The small margins in the predicted outcomes emphasize the potential volatility of the race; a slight shift in voter sentiment or late-breaking events could easily flip the outcome in these battleground states. This narrow margin of victory suggests a level of electoral fragility for Harris, as her support base may not be solidly locked in. The close race in each state implies that both candidates must maintain their voter engagement and strategic focus, as any shift in these states could have significant implications for the election's overall outcome. Thus, while Harris is favored in the majority of battleground states, the slight leads indicate a need for cautious interpretation, with both campaigns facing high stakes in these key states.

5.2 Real World Implications

If Kamala Harris wins the election by a narrow margin, the results could have significant real-world implications, reflecting the deeply polarized state of the nation. A close race indicates not only a tight political divide but also underscores broader cultural, economic, and social differences among Americans, often manifested through political affiliations. One major challenge for Harris would be governing with such a divided mandate; she could face resistance from nearly half the population, complicating her efforts to implement her agenda. This suggests that any significant policy shifts may encounter strong opposition, potentially deepening divisions rather than bridging them [@.

Moreover, a narrow victory for Harris would reinforce national polarization, revealing the extent to which the country is split. The nearly equal support for both candidates highlights entrenched ideological divides, with voters committed to opposing perspectives on critical issues such as healthcare, immigration, climate policy, and the role of government (Heltzel & Laurin, 2020). Economic concerns, including job security, wage equality, and tax policy, alongside cultural values related to family, education, and identity, play significant roles in voter alignment. Thus, a close election result suggests that these differences are sharply felt

across the electorate, with each side viewing the outcome as an opportunity to uphold their values and address their unique challenges.

5.3 Uncertainty Beyond Data and Modeling

There are external factors other than data and modeling that could influence the actual outcome of the election. These factors are often hard to control and predict.

Firstly, recency bias influences public perception and even polling responses. Many people expect Trump to win because of his unexpected victory in 2016 and his close race in 2020, despite lagging in the polls. This bias conveys the tendency to focus on recent, high-profile events while overlooking historical trends, such as Obama's strong polling performance in 2012, where he outperformed polls. This bias can affect how people respond to polls or even how pollsters interpret results, as certain events may be weighted more heavily than warranted. As our model depends on polling data, any systematic bias introduced by recency expectations can skew our predictions, making it challenging to account for potential polling errors.

Social desirability bias and non-response bias are additional sources of uncertainty, particularly with Trump's supporters. Some voters may be hesitant to disclose their support for a conservative candidate due to social stigma, leading to underreporting of support in polls. This reluctance can be especially strong in areas where conservative views are less socially accepted. Trump's support base also includes individuals with lower social trust and civic engagement, which means they are less likely to participate in polls altogether. If certain groups systematically avoid participating, polling data may underrepresent these voters, resulting in an underestimation of Trump's support in our model. This non-response bias is difficult to quantify, adding a layer of uncertainty to our predictions. On the other hand, Harris faces potential biases such as the Bradley effect and gender-related bias. As a Black female candidate, Harris might encounter the Bradley effect, where some voters express indecision rather than explicitly stating they won't support a minority candidate. Historically, this effect has impacted Black candidates, as undecided voters tend to lean against them in the actual vote count, despite polling responses suggesting neutrality. Additionally, Clinton's 2016 campaign demonstrated that undecided voters can heavily swing against female candidates. This trend could similarly impact Harris's polling and election results. If these voters decide against Harris in the end, our model might overestimate her support based on the polling data, leading to inaccuracies in predictions. These biases are inherently difficult to measure or adjust for, as they reflect deeper societal attitudes and individual decision-making processes. Pollsters may attempt to correct for non-response or social desirability bias, but such adjustments are often based on assumptions that may not fully capture the complex social dynamics at play. Consequently, our model's predictions should be interpreted with caution, acknowledging that these biases can lead to unexpected shifts in voter behavior that fall outside the range of our data-driven forecasts.

5.4 Limitations and Weaknesses

Although the model is suitable and performs well on achieving the goal of predicting the outcome, it has several limitations that could be addressed in future studies. Here are some key areas:

- **Pollster Ratings and Credibility**: Polls from different organizations often vary in reliability, as some pollsters have more accurate or less biased methodologies. By integrating pollster ratings, the model could assign weights to different polls based on the historical accuracy and reputation of each organization. This adjustment would allow the model to downplay the influence of potentially biased or less credible polls, particularly in competitive battleground states, where small differences in predicted support can significantly affect the outcome. This approach would lead to a more calibrated model that accounts for poll variability and thus provides a more reliable prediction.
- **Timing and Recency of Polls**: Voter sentiment is not static, and the timing of polls can significantly impact their relevance. As the election date nears, polls generally become more reflective of the final outcome. Incorporating a temporal weighting scheme that prioritizes recent polls could enhance the model's responsiveness to last-minute shifts in voter preferences, which often occur due to campaign events, debates, or emergent political issues. This weighting approach could also help mitigate the risk of using outdated data that no longer accurately reflects current voter sentiment.
- **Economic and Political Context**: The political landscape is shaped by various external factors, such as economic indicators (e.g., unemployment rates, inflation) and prominent political events. By incorporating economic data or major political issues, such as health care or immigration policies, the model could capture contextual variables that influence voter sentiment. Such contextual factors could be especially useful in a Bayesian framework, allowing us to adjust prior distributions based on known historical influences on voter behavior.
- **Inclusive of National/Popular votes**: our model has exclusive focus on the seven battleground states—Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. By narrowing the dataset to only these critical swing states, our model omits data from the rest of the nation, including states with predictable voting patterns (solidly Democratic or Republican) and states outside the battleground category. This omission is significant as it restricts the model's capacity to capture the broader national voting landscape and limits the ability to provide a holistic prediction of the 2024 U.S. Presidential Election outcome. In particular, the exclusion of national popular vote data limits our model's understanding of overall voter sentiment across the country. While battleground states are pivotal in determining the electoral outcome due to the Electoral College system, popular vote trends from non-battleground states can still reveal important national shifts in sentiment that may indirectly influence battleground dynamics. For example, a nationwide trend towards or away from a candidate could suggest emerging political momentum or reflect the impact of major events that also resonate within battleground states. Without incorporating this popular vote data,

our model may overlook broader trends that could subtly shape battleground results, especially if voters in battleground states are influenced by national sentiment.

5.5 Implications for Future Modeling

For future modeling, incorporating data from a broader range of states or using a weighted approach that balances battleground states with national polling data could address these limitations. For example, incorporating popular vote trends and data from states with similar demographics or political contexts could help the model detect and account for national sentiment shifts. Additionally, adding more factors, such as ratings for pollsters, demographic details, and economic conditions, could help the model capture different influences on how people vote. Giving more weight to recent polls would also make the model more responsive to last-minute changes in public opinion as Election Day gets closer. By expanding the data and including a wider range of details, future models could better balance the unique characteristics of battleground states with trends across the nation, resulting in a stronger and more reliable prediction tool for elections.

Appendix

A Additional Data Details

A.1 Data Cleaning

The data cleaning process for our analysis focused on refining the raw polling data obtained from Project 538's polling project (FiveThirtyEight 2024). The raw data contains polling information for various U.S. states, but for our specific analysis, we concentrated on seven designated battleground states: Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. This is due to the fact that these battleground states have been recognized as pivotal in determining this year's election outcome.

To prepare the data for modeling, we performed several key steps. First, we filtered the data to retain only relevant columns, including state, polling dates, pollster information, numeric grades, transparency scores, candidate names, and their respective support percentages. This filtering process ensured that we excluded any unnecessary data that could complicate our analysis. Next, we implemented thresholds for the numeric grade (minimum of 2.5) and transparency score (greater than 5) to ensure that only high-quality polling data was considered. We also converted polling dates into a standardized date format and created a new numeric variable representing the number of days since the earliest polling date. Additionally, we handled specific data inconsistencies, such as replacing "Nebraska CD-2" with "Nebraska." The polling percentages were then aggregated by averaging the values for each candidate while preserving other important attributes such as pollster and state. This aggregation allowed us to focus on polling data that represented a balanced view of voter support across the selected battleground states. The cleaned dataset was then saved in a parquet format for efficient processing in subsequent analyses.

To ensure the integrity of our cleaned polling dataset, we implemented a comprehensive data testing process. This included verifying the presence of all required columns, checking data types for consistency, and ensuring there were no missing values in critical fields. We validated that candidate support percentages fell within the logical range of 0 to 100 and confirmed that numeric grades and transparency scores met predefined thresholds. By rigorously applying these tests, we ensured that the dataset is reliable and suitable for subsequent modeling and analysis, providing a solid foundation for our exploration of voter behavior in the battleground states.

A.2 Limitations and Future Directions

To improve data selection for future studies, data from non-battleground states could add valuable context and enhance model robustness by reflecting a range of voter behaviors and regional influences. While these states may not directly impact the electoral outcome due to

their predictability, understanding regional dynamics across all states could help the model detect shifts in voter sentiment that cut across state lines. For instance, trends in nearby states or those with similar demographic or economic characteristics might mirror those in battleground states, providing additional insights into potential voting patterns. Including these states in the model could also reveal hidden correlations and strengthen the model's predictive power by capturing a wider range of political and demographic factors. The focus on battleground states also means that the model may amplify the idiosyncrasies of these seven states without balancing them against broader national trends. Each battleground state has unique political, economic, and demographic characteristics, and relying solely on these states could cause the model to disproportionately weigh regional dynamics that may not be representative of the country as a whole. As a result, predictions based only on these states might overestimate or underestimate certain voting behaviors, particularly if the election outcome depends on trends that are not captured within these seven states alone.

B Additional Model details

B.1 Model Justification - Further Details

The Bayesian Beta Regression model with a logit link function is particularly suited for our prediction of the 2024 U.S. Presidential Election due to its alignment with the nature of our outcome data and the goals of our analysis. Our primary outcome variable is the percentage of support for each candidate in the seven battleground states, which we scaled to a proportion between 0 and 1. Since support percentages are naturally bounded between 0% and 100%, modeling these data with traditional linear regression presents challenges, as linear regression assumes an unbounded continuous outcome. The Beta distribution is ideal for this type of data because it is defined on the interval $(0, 1)$, allowing us to model proportions without the risk of generating unrealistic values outside this range. Unlike a linear model, which could predict support percentages exceeding 100% or falling below 0%, the Beta distribution respects the boundaries of our data.

To effectively capture the dynamics of voter support, we utilized natural splines with the variable end_date_num, which represents the number of days since the earliest poll date. This choice allows the model to adapt to non-linear trends in voter support that can fluctuate due to various campaign events, debates, or shifts in the political landscape. By specifying four degrees of freedom for the natural splines, we enable the model to flexibly capture gradual changes in voter sentiment while mitigating the risk of overfitting to noise or abrupt fluctuations in the data. This flexibility is particularly critical in electoral polling, where opinions may shift slowly or rapidly in response to unpredictable external events. Alternative models, such as standard logistic regression without splines, would fail to capture these complex changes, potentially leading to a less accurate depiction of how support varies over time.

A unique aspect of our analysis is the focus on battleground states, which exhibit distinct political dynamics that can significantly influence election outcomes. Incorporating state as a random effect provides the flexibility necessary to estimate state-specific baseline support levels for each candidate while still allowing us to identify overall trends across all states. This approach is more effective than treating state as a fixed effect, as it avoids overparameterizing the model with individual state coefficients, which could lead to overfitting, particularly in cases with limited polling data for certain states. The random effect structure enhances the model's generalizability, making it more robust when predicting support in battleground states with varying polling frequencies.

The Bayesian approach offers substantial advantages for our model by allowing us to incorporate prior information, which enhances the stability and reliability of our estimates. Given that polling data can be sparse for certain states or dates, Bayesian inference helps stabilize parameter estimates, especially when limited data points are available. We defined weakly informative priors for the model parameters: for the intercept β_0, we specified $\beta_0 \sim \text{Normal}(0, 10)$ to provide a relatively uninformative prior that allows the data to inform the estimation of baseline log-odds; for the slope β_1, we used $\beta_1 \sim \text{Normal}(0, 5)$, reflecting reasonable expectations regarding the influence of polling trends over time.

This model structure not only allows for capturing the dynamics of voter support but also provides interpretability that aligns well with the goals of our paper. The logit link function enables us to interpret results in terms of odds or probabilities of victory, which is crucial in the election context. By constructing separate models for Trump and Harris, we can directly compare their temporal trends and state-level variations without confounding effects. Thus, the Bayesian Beta regression model with a logit link function is justified as the most suitable model for our study. It accommodates flexible non-linear trends over time, accounts for regional variation across battleground states, and delivers interpretable results concerning victory probabilities. The Bayesian framework not only stabilizes estimates in the presence of sparse data but also provides meaningful uncertainty estimates, ultimately aligning with our objective of forecasting each candidate's probability of victory in the 2024 U.S. Presidential Election.

B.2 Model Diagnostics

To assess the validity and reliability of our Bayesian models for predicting voter support for Kamala Harris and Donald Trump, we conducted posterior predictive checks for both models. These checks are essential for evaluating the models' ability to replicate observed polling data based on the estimated parameters. By comparing the distributions of observed support percentages with the predicted values, we can identify discrepancies and assess the overall fit of our models.

Posterior predictive checks help verify model assumptions and provide insights into how well the models capture the underlying voter dynamics. A close alignment between observed and

predicted values suggests that our models effectively reflect voter support trends in the electoral context. In contrast, significant deviations may reveal areas for model improvement or highlight limitations in our approach. This diagnostic process aims to enhance the credibility of our findings and deepen our understanding of the electoral landscape as the 2024 Presidential Election approaches

B.3 Posterior Predictive Check for the Trump Model

The posterior predictive check for the Trump model is illustrated Figure 10, showcasing the relationship between the observed support percentages y and the predicted support values y_{rep}. The plot displays a density estimate for both the actual observed data (in dark blue) and the predicted values generated by the model (in light blue). The primary aim of this diagnostic check is to assess how well the model captures the distribution of observed support for Donald Trump.
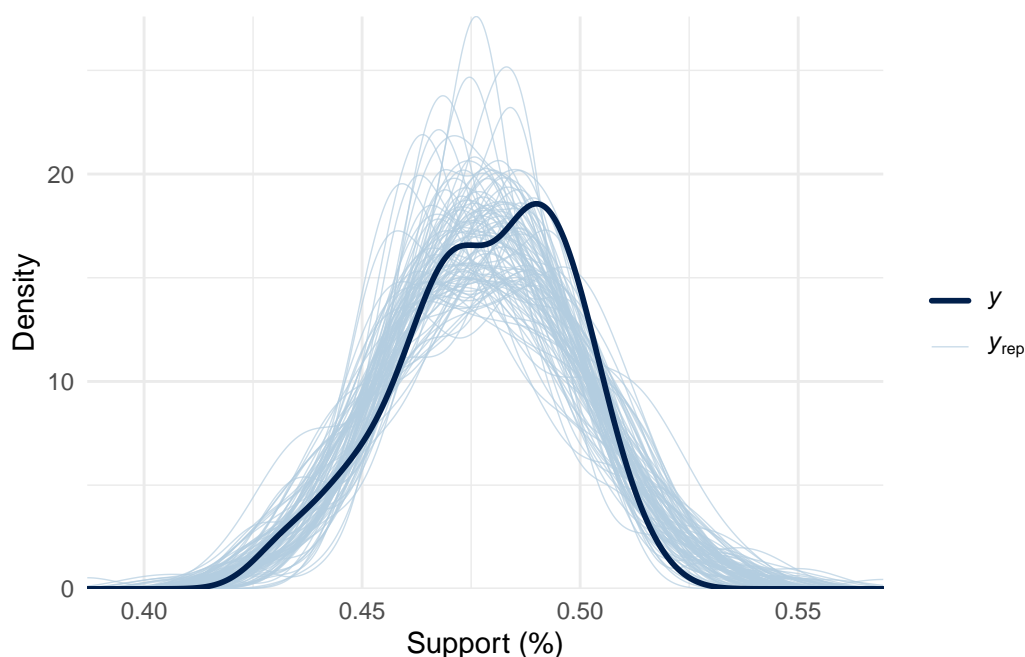

Figure 10: Posterior predictive check for Donald Trump's model, illustrating the distribution of observed support percentages y alongside the predicted support values y_{extrep}. The graph shows a strong overlap between the observed and predicted distributions, indicating that the model accurately captures the dynamics of voter support for Trump.

From the visualization, we can observe that the predicted values closely align with the actual observed support percentages. The dark blue curve representing the observed data indicates a

concentrated support range around 48% to 52%, suggesting that the model effectively captures the central tendency of voter support for Trump. The light blue curves depict a variety of predicted densities, indicating the model's ability to simulate a range of possible outcomes around the observed values. The alignment of the y and y_{rep} distributions indicates that the model is successfully reflecting the dynamics of voter support, thereby validating the choice of model and its structure. However, the presence of some variability in the predicted values suggests that while the model fits the observed data well, there may still be unexplained fluctuations in support levels that could warrant further investigation. These limitations are further discussed in Section B.2. This examination enhances our confidence in the model's utility for forecasting electoral outcomes, ensuring it is robust enough to inform strategic decisions as the election approaches.

B.3.1 Model Diagnostics - Harris Model

The posterior predictive check for the Harris model illustrates the relationship between the observed support percentages y and the predicted support values y_{rep}. Figure 11 displays a density estimate for both the actual observed data (depicted in dark blue) and the predicted values generated by the model (shown in light blue). The purpose of this diagnostic check is to evaluate how well the model captures the distribution of observed support for Kamala Harris.

In the plot, we observe a strong overlap between the y and y_{rep} distributions, with the dark blue curve representing the observed support percentages primarily concentrated around 48% to 52%. This indicates that the model successfully approximates the actual voting support dynamics for Harris. The light blue curves signify the distribution of predicted values, demonstrating the model's ability to replicate the observed trends effectively. The close alignment of the observed and predicted densities suggests that the model adequately reflects the underlying support for Harris among voters, thereby affirming the appropriateness of the model structure. However, similar to the Trump model, the presence of some variability in the predicted values indicates that while the model performs well, there may still be unexplained variations in support levels that could benefit from further analysis. Such limitations are also discussed in Section B.2.

B.3.2 Model Performance

Based on the posterior predictive checks performed for both Trump and Harris models, we are able to make the following conclusions on the performance and reliability of our Bayesian regression models. The strong overlap of the observed distribution of support percentages and the predicted distribution from posterior samples suggests that the models are well-calibrated, meaning they accurately reflect the variability and central tendency of voter support for Trump and Harris, respectively. This alignment is particularly important given the fluctuating nature
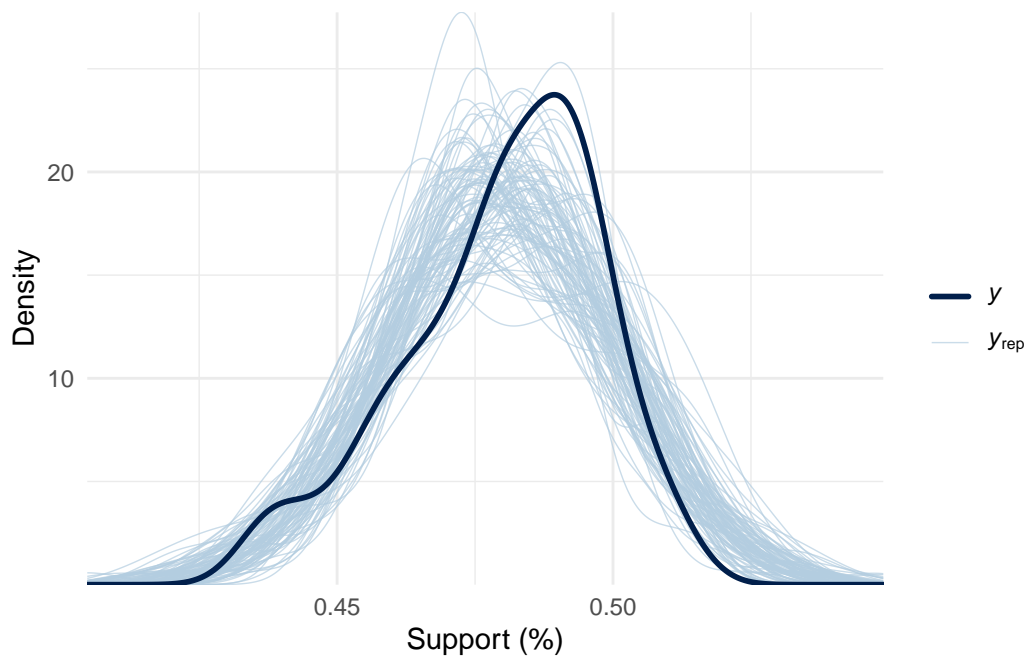
Figure 11: Posterior predictive check for Kamala Harris's model, illustrating the distribution of observed support percentages y against the predicted support values y_{extrep}. The graph shows a strong correspondence between the observed and predicted distributions, indicating that the model accurately captures the dynamics of voter support for Harris in the given dataset.

of polling data and the importance of accurately capturing voter sentiment as we approach Election Day.

In the context of our paper, these predictive checks provide evidence of the model's robustness and reliability in forecasting. Given our goal of predicting the likelihood of each candidate's victory in key battleground states, it is important that our models can not only predict the general trend in support but also approximate the changes in voter sentiment captured by the polling data. The successful predictive checks validate our choice of a Bayesian Beta regression model with a logit link function, demonstrating that this approach is indeed suitable for analyzing proportion data constrained within the 0-1 range.

Furthermore, these results support the use of state-level random effects and time-based splines, as the models are shown to accommodate both state-specific variations and temporal shifts in support. By accurately modeling these elements, our predictions provide more reliable insights into which candidate holds an advantage in each state. This model performance enhances confidence in our forecasts, suggesting that our predictions can serve as a meaningful tool for understanding electoral dynamics in these critical regions. These findings suggest the model's capacity to provide robust, data-driven projections of the election outcome, while also accounting for the inherent uncertainty and volatility within polling data.

C Pollster methodology overview and evaluation: NYT/Siena College

C.1 Background of NYT/Siena College Polling

The NYT/Siena College Polling is a U.S. polling organization where the New York Times and Siena College collaborate to deliver precise and timely poll results for the presidential elections (The New York Times 2023). For the 2024 election cycle, The NYT/Siena College poll uses a robust sampling approach to ensure accuracy and representativeness. The poll applies a stratified dual-frame sample, drawing from both landlines and cell phones.

C.2 Target population, Sampling frame, and Sample

- Target population: Target population refers to "the collection of all items about which we would like to speak" (Rohan Alexander 2024). In this case, NYT/Siena College's polling target population is all registered American voters who live in the six battleground states: Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin. Given the nature of the presidential elections, the election result is based on the electoral college instead of the popular vote, thus the pollster focuses on polling on the states that are the likeliest to decide the outcome of the race, which are called battleground states. The battleground states are states that have similar voter support for both the Democratic

or the Republican party, a slight difference in support could determine which party wins a state.

- Frame: Sampling frame refers to "a list of all items from the target population that we could get data about" (Rohan Alexander 2024). NYT/Siena College's sampling frame is a comprehensive list of registered American voters obtained from the L-2 voter file, it includes details such as demographic characteristics, contact information, and voting history. By using the L-2 voter file as their sampling frame, the pollster can ensure representativeness by allowing them to draw samples that accurately reflect the demographics of the registered voter population.
- Samples: Sample refers to "the items from the sampling frame that we get data about" (Rohan Alexander 2024). In this case, the samples of NYT/Sienna College are individuals who are registered American voters obtained from the L-2 voter file from each stratum from each state who have also answered and completed the questionnaires.

C.3 Sampling Methodology

NYT/Siena College uses stratified sampling as their methodology where the population is divided into distinct strata based on shared characteristics. These strata include age, gender, race, geographic region, and other various relevant variables. Once the population is divided, random samples are taken from each stratum, in proportion to their size in the population. For each poll, NYT/Siena College takes a sample size of 1000 registered American voters. To refine their sample, NYT/Siena College adjusts the data by key demographic and political variables such as region, race/ethnicity, party affiliation, education, and voting patterns from the 2020 election. These adjustments help ensure the sample accurately represents the electorate. Additionally, they use a "likely voter screen," which combines self-reported likelihood of voting with historical voter behavior to estimate how likely respondents are to vote in the upcoming election. At every step of the survey, Siena/NYT uses the information in the datasets to try to ensure that they have the right number of Democrats and Republicans, young people and old people, the right ratio of people with different income levels, and a diverse mix of different races and regions. Once the survey is complete, they compare their respondents to the voter file and use a process known as weighting to ensure that the sample reflects the broader voting population. This combination of historical data and weighted adjustments ensures the poll is designed to predict election outcomes as accurately as possible.

C.4 Trade-off of Sampling Methodology

Although stratified sampling has many strengths such as appropriate representation of the target population, reduction of sampling bias, and sampling efficiency, it also has some weaknesses. First, its implementation can be complex, stratified sampling requires detailed knowledge of the population to define the strata appropriately. This can be challenging to implement if accurate, up-to-date information on demographic or geographic characteristics is unavailable

or difficult to access. For example, if there are significant changes in the L-2 voter file since the last election, it is difficult to take account into the unknown changes. Second, there could be potential for mis-stratification. If the chosen strata do not capture meaningful differences within the population, or if important subgroups are overlooked, the results might be biased or inaccurate. For example, if new voting trends emerge that were not captured in the strata definitions, the poll might miss important shifts in voter opinion. Third, stratified sampling can be time-consuming and resource-intensive. Diving the population into strata and ensuring proportionate sampling within each group can be more time-consuming and resource-intensive compared to simple random sampling. Pollsters may need to gather detailed data on the target population, which adds to the cost and complexity of the survey. Lastly, there is the risk of over-emphasis on stratification. If the population is stratified too finely, the sample size in each stratum may become too small to yield meaningful results, increasing the margin of error for each subgroup.

C.5 Sample Recruitment Methods

Each poll is conducted by phone using live interviewers at call centers based in Florida, New York, South Carolina, Texas, and Virginia. Most people that the live interviewers call do not answer the call, and some don't finish the survey, leaving only about two percent both answer their phone and complete the poll. To mitigate this: First, the pollster keeps calling numbers until someone picks up, even though it can take hours–a reason why a high-quality poll is time-consuming and expensive to conduct. Second, voter files are used to help make sure the pollster is reaching a representative sample; for example, they may reach out to a higher percentage of people who are less likely to respond. Third, pollster mathematically adjusts the responses they do receive so they are representative of the wider population, it includes adjusting by age, race, gender, voting history and etc.

C.6 Non-response Handling

Non-responses are critical factors in polling because they affect the representativeness of the sample and, consequently, the accuracy of the poll results. When a significant portion of selected individuals doesn't respond, the poll risks becoming unrepresentative of the broader population, especially if certain groups are systematically more or less likely to respond. It is also important to know how pollsters handle non-response since it provides transparency about the polling methodology. It also gives confidence that the results are not skewed by an unrepresentative sample when pollsters clearly explain their non-response handling. Therefore, polls with thorough non-response strategies are generally more reliable than those without. The NYT/Siena College pollsters handle non-response by using weighting and adjusting their sample to correct for any biases that might emerge due to individuals not responding. Specifically, they adjust on multiple demographic and political variables such as age, gender, education, race/ethnicity, and party affiliation. They also account for variations in voter likelihood and

previous voting patterns, ensuring that the sample represents the likely electorate as accurately as possible. This helps reduce the potential bias from non-response, especially since certain demographic groups or political affiliations may be less likely to respond to polls. By rebalancing the sample, they ensure that even if some groups have lower response rates, their representation in the final poll results aligns with what is expected based on historical trends and current voter enthusiasm.

C.7 Questionnaire Pros and Cons

Based on the NYT/Siena College's past practices on presidential election polls, they often ask questions about candidate preference, approval ratings, voter opinions on key issues, economic perceptions, as well as social and cultural issues. They also ask about the respondents' demographic information and voting behavior to categorize respondents for analysis.

C.7.1 Strengths of the NYT/Siena College questionnaires

First, they have a comprehensive question design. NYT/Sienna College polls ask a wide range of questions that provides a detailed view of the respondent's preferences and opinions. It helps produce multi-dimensional data that allows for in-depth analysis and segmentation of voter groups. Second, the questionnaires are tailored to likely voters, as identified through the L-2 voter file. This means that the voters are more likely to be up-to-date with relevant policies proposed by different parties, and have formed their own opinions about them, this improves the predictive reliability of the poll. Third, the questionnaires contain detailed demographic questions about the respondents which allow NYT/Siena College to stratify responses by voter subgroups, identifying varied voter bases and diverse issues.

C.7.2 Weaknesses of the NYT/Siena College questionnaires

First, the length and complexity of the questionnaires might cause the respondents to be less likely to respond or finish the polls. Respondents may rush through questions or drop out before completion. Respondents may also be prone to moderacy response bias where they have the tendency to choose middle options regardless of question content, as well as response order bias where respondents choose answers based on their order. Second, questionnaires have a dependence on self-reported likelihood to vote. Self-reported likelihood can be unreliable, especially for individuals whose intentions to vote may change close to the actual election date. This can reduce the accuracy of predictions based on likely voter models. Third, there are limited open-ended questions. Many voters may have mixed opinions on specific issues or policies that cannot be explained in a single-choice answer. However, answers to open-ended questions are often difficult to convert to actionable data, so questionnaires tend to rely on closed-ended questions for scalability. This might restrict the ability to capture complex

reasonings behind their choices, which could add extra strain on the time and resources used to interpret the answers.

D Idealized Methodology

Building on our discussion of the New York Times/Siena College Poll, we now present an idealized methodology that aims to enhance the accuracy and reliability of our predictions. This approach incorporates best practices and innovative techniques to ensure a comprehensive prediction of voter behavior in the upcoming 2024 U.S. presidential election. To demonstrate our idealized methodology, we generated a survey on the 2024 US Presidential election, which can be accessed via the the URL provided in (Google 2024) under Section D.2.

D.1 Sampling Approach

Using data from voter files, which contain demographic information about registered voters, we aim to ensure that our sample accurately represents the population. To achieve this, we will use stratified random sampling. Given the importance of state-level forecasts due to the Electoral College system, we will first stratify by state.

Within each state, we will apply stratified sampling again, dividing the population into 6 groups and selecting 100 random samples from each stratum within those groups, ensuring our sample captures key demographic critical for predicting voter behavior. The groups and their respective strata include age groups (18-29, 30-44, 45-64, and 65+), gender (Male, Female, and Other), and race/ethnicity (White, Black, Latino, Asian, and Other). Additionally, household status will be categorized as either renting or owning, with home ownership serving as a proxy for wealth. Voting history will be classified into those who voted in the previous election and non-voters, while party registration will include Democrats, Republicans, and Independents.

For the recruitment of participants, we will begin by sending survey invitations to our selected voters via email, offering a $20 incentive to encourage participation. To improve response rates and reduce non-response bias, we will send a reminder email three days later. This approach will help minimize expenses to some extent. If we do not receive a response within a week (7 days), we will follow up by phone to reach those who did not respond to the email invitation.

D.2 Data Validation

After completing the survey, we will adjust the data through weighting to ensure the sample accurately reflects the broader population for predictive purposes. More weight will be assigned proportionally to each stratum based on its size, and additional weight will be given to respondents from strata that are less likely to participate in surveys. And to avoid duplicate responses, each voter will be assigned a unique ID with only the first response from

each ID being retained. To ensure the selected voters are completing the survey correctly, we will track responses in real-time using a centralized system. Additionally, we will include validation questions in the survey to catch careless or fraudulent responses.

References

Albers, L. E. et al. 2023. *Usmap: US Maps Including Alaska and Hawaii.* https://CRAN.R-project.org/package=usmap.

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary Tables and Plots for Statistical Models and Data.* https://CRAN.R-project.org/package=modelsummary.

Bates, Douglas, and Martin Maechler. 2023. *Matrix: Sparse and Dense Matrix Classes and Methods.* https://CRAN.R-project.org/package=Matrix.

Bolker, Ben et al. 2023. *broom.mixed: Tidying Methods for Mixed Models.* https://CRAN.R-project.org/package=broom.mixed.

Bürkner, Paul-Christian. 2023. *brms: Bayesian Regression Models using 'Stan'.* https://CRAN.R-project.org/package=brms.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. *FiveThirtyEight: 2024 US Presidential Election Polls.* https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Gabry, Jonah et al. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan.* https://mc-stan.org/rstanarm/.

Google. 2024. *Google Form: [2024 US Presidential Election Survey].* https://docs.google.com/forms/d/1S6aN5Q82orjKYJI4i7001sPyn5jhRkgNIkI9teiQq_o/edit.

Grolemund, Garrett, and Hadley Wickham. 2011. *Lubridate: Make Dealing with Dates a Little Easier.* https://CRAN.R-project.org/package=lubridate.

Kuhn, Max et al. 2023. *Caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal et al. 2023. *Arrow: Integration to Apache Arrow for Data Frames.* https://CRAN.R-project.org/package=arrow.

Rohan Alexander. 2024. "Telling Stories with Data." https://tellingstorieswithdata.com/.

Team, R Core. 2023. *splines: Regression Spline Functions and Classes.* https://stat.ethz.ch/R-manual/R-devel/library/splines/html/00Index.html.

The New York Times. 2023. "How the Times and Siena College Poll Was Conducted." https://www.nytimes.com/article/times-siena-poll-methodology.html.

Wickham, Hadley et al. 2023. *Tidyverse: Easily Install and Load the 'Tidyverse'.* https://CRAN.R-project.org/package=tidyverse.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.