

My title*

My subtitle if needed

Chenming Zhao

October 31, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

2 Data

2.1 Overview

Polling data for the 2024 U.S. presidential election is provided by FiveThirtyEight (FiveThirtyEight 2024). This dataset compiles information from numerous polls conducted by various pollsters, detailing public support levels for presidential candidates. This data is periodically updated on FiveThirtyEight’s website to reflect the latest polling information for the 2024 election. Each record includes information we need like state in which the poll was conducted (if applicable), polling dates (start and end), the specific pollster, the sample size of each poll, the polling methodology (such as phone, online, etc.), and support percentages for each

*Code and data are available at: https://github.com/RohanAlexander/starter_folder. We gratefully acknowledge FiveThirtyEight for providing the polling data used in this analysis. The data is available at: <https://projects.fivethirtyeight.com/polls/president-general/2024/national/>.

candidate. This dataset serves as the foundation for analyzing trends and building predictive models for the winner of the upcoming US presidential election. Table `tbl-pollingdata` below previews this information for polls conducted in 2023 and 2024.

To simulate, test, analyze, and clean the polling data, the statistical programming language R was employed (R Core Team 2023). Key libraries that supported the data analysis include `tidyverse` (Wickham et al. 2019), `readxl` (Wickham and Bryan 2023), `dplyr` (Wickham et al. 2023), `knitr` (Xie 2023), `ggplot2` (Wickham 2016), and `broom` (Robinson, Hayes, and Couch 2024) for tidying model outputs. Additionally, the `rstanarm` (Goodrich et al. 2022) library was utilized to implement multilevel modeling for prediction purposes. Also this analysis inherits the folder structure from Professor Rohan Alexander from University of Toronto (Alexander 2024).

2.2 Measurement

The process of transforming real-world phenomena into structured data typically involves several critical steps, from designing surveys to final data entry. First, we identify what needs to be measured. In this case, we focus on public opinion regarding the 2024 presidential election, particularly support rates for various candidates. It is essential to define the data scope (e.g., national or state level) and key variables—such as polling information (e.g., polling organization, start and end dates), sample details (e.g., sample size and demographics), and candidate details (e.g., candidate name and party affiliation).

Next, polling organizations develop survey questions based on the desired data, considering sample size, sampling methods, and survey modes to capture respondents' preferences and voting intentions. Different methodologies impact the results, as they reflect the views of distinct groups, which may be influenced by factors like access channels, demographics, and technology usage. For instance, face-to-face surveys tend to capture opinions from offline-inclined groups, whereas voluntary online surveys exclude those less active online cite .

Once the survey is designed, polling begins, and responses are aggregated by the polling organization. Respondents' answers are recorded and categorized, converting their opinions into structured data points, though these lack the nuances of human complexity. To protect privacy, data are often anonymized and aggregated, retaining key insights. After collection, raw data undergo cleaning and processing to minimize sampling biases and errors. Yet, any single poll can have random and systematic error sources, so combining multiple poll results helps mitigate these errors cite .

Processed data are entered into structured datasets. Ideally, well-formatted polling data are stored as individual records, typically including fields like polling organization, sample size, survey dates, margin of error, candidate support percentages, and methodology description. On platforms like FiveThirtyEight, these fields are standardized for ease of comparison and analysis. Finally, FiveThirtyEight and similar platforms aggregate these structured records

and present them to the public or allow non-commercial research use, with data usually updated as new polls are released. This enables analysts and the public to track changes in candidate support, analyze trends, and interpret results. An interesting phenomenon, however, is that as election day approaches, poll results often “converge” or “follow” the average, reducing polling error, with evidence suggesting some “final adjustments” in methodology may occur cite .

2.3 Analysis Data

This analysis is focused on Donald Trump, so we only collected observations where the `candidate_name` is Trump. At the same time, we prioritize the quality of the polls. In a political approval rate prediction model, which is often influenced by numerous complex factors, we encounter data that is typically complex and noisy. Low-quality poll results can introduce significant bias into our prediction model, further impacting the accuracy of inferences. Therefore, only polls with a `numeric_grade` above 2.7 were selected. Additionally, Trump’s main opponent, Kamala Harris, only entered the race after Biden’s withdrawal, a major political event that could have significant effects on the support rates for both sides. As such, we focused solely on polls conducted after Biden’s withdrawal to enhance the model’s prediction accuracy for the final outcome.

In researching the likelihood of Trump winning, we selected specific variables to focus on those most impactful for predicting approval rates. Here is an overview of data used for this analysis Table 1. Following are selected outcome variables and predictor variables, and explanations respectively.

2.3.1 Outcome variables

In this analysis of Donald Trump’s support rate in the 2024 U.S. presidential election, the outcome variable is Trump’s support rate (pct) because it is directly related to the election results. The primary research goal of this analysis is to predict Trump’s chances of winning in the 2024 election, so the support rate is the main metric we aim to model and forecast. The support rate represents the proportion of respondents in a specific poll who support Trump. We seek to understand how the support rate is influenced by different factors, such as polling methodology, time, or geographical variables. By modeling the fluctuations in Trump’s support rate, we can infer his electoral performance, analyze the trends within his support base, and assess the impact of various factors on his popularity. Understanding these dynamics provides insights into Trump’s standing relative to his opponents and supports our overall election prediction.

2.3.2 Predictor variables

In this analysis, the predictor variables are the factors that we believe can influence the support rate for Trump. These variables are chosen based on their potential to explain changes in support rate. Following are the predictor variables that will be included in our model, along with an explanation of their significance.

start_date: Public sentiment and approval rates fluctuate dynamically during an election period, influenced by the passage of time, major events, and campaign strategies. By recording the date of each poll, we can observe changes in support over time.

state/national: Political environments and voter preferences may vary significantly by state, and national polls cannot fully reflect conditions in individual states. The electoral college system in the U.S. makes state-level approval rates crucial for predicting the likelihood of winning.

pollster: Different polling organizations use varying survey methods, sampling strategies, and data-processing procedures, which impact the reliability and accuracy of the poll results. Knowing which polling organization published the data helps assess potential biases in the data.

methodology: Different survey methods (e.g., telephone interviews, online questionnaires) significantly impact the representativeness of the results, as they directly influence the characteristics of the sampled population. Some methods may introduce biases toward specific demographic groups; for example, telephone surveys may skew toward older demographics, while online surveys may skew toward younger ones.

is_national: This variable helps us distinguish between national and state-level polls, ensuring the model can adapt and adjust across different geographic levels.

days_since_Biden-Withdrawal: Trump’s main opponent, Kamala Harris, entered the race only after Biden withdrew. Major political events like Biden’s withdrawal can significantly impact voter approval rates, and tracking the time since this event helps assess both its short-term and long-term effects.

Other potential variables (such as voter demographics like gender, age, race, and income) could indeed influence voter approval rates, but they primarily pertain to a **micro-level** analysis. After incorporating GDP and per capita income data from various U.S. states and nationwide in 2023 into a generalized linear model, we found that their correlation with presidential election outcomes was not very strong, likely due to data limitations in capturing the effect of economic changes on the election. Thus, we did not include these factors. Additionally, individual voter characteristics (such as gender and education level) are not consistently available in most polls. In comparison, the variables we chose are more complete and comprehensively capture the macro nature of the data. Therefore, selecting these variables allows us to maintain data simplicity while maximizing the capture of key factors influencing approval rates, avoiding the noise introduced by excessive micro-level variables.

Table 1: The First Few Lines of Cleaned Analysis Data

start_date	state/national	pollster	methodology	Trump Support Rate	is_national	days_since_Biden_Withdrawal
2024-10-23	Texas	Siena/NYT	Live Phone	52	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	50	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	52	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	51	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	55	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	53	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	55	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	53	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	42	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	41	0	94

2.3.3 Figures

The following figures, Figure 1, Figure 2 and Figure 3, display various aspects of data related to Trump’s support rate, helping us understand how support rates vary with different factors. In Figure 1 and Figure 2, we have marked the occurrence count of each pollster or state on the x-axis, helping us identify potential sources of bias in the data. Low-frequency items (such as data points from certain pollsters or states) may reduce the representativeness of the model—low-frequency observations imply insufficient data for specific pollsters or states, which may not accurately reflect voter support rates in these areas. This incomplete data could lead to unreliable predictions for these pollsters or states, increasing prediction errors. Additionally, low-frequency data is more susceptible to extreme values. With fewer observations, certain outlier data points (e.g., a particularly high or low support rate in a single poll) can appear disproportionately significant, affecting the overall prediction. When fitting the model to low-frequency data, it may be more influenced by the “noise” in these data points, potentially introducing bias into the overall results. Systematic biases in low-frequency observations could spread throughout the model, impacting the overall accuracy of the predictions.

Although low-frequency observations may introduce some bias, directly removing them would result in information loss, especially in election predictions where the representativeness of different states and pollsters is crucial. Different pollsters and states have their own resources and priorities, leading to uneven data coverage. Some pollsters may prefer to publish data at specific times, while certain states, due to smaller populations or stable voter behavior, may receive less attention. Thus, it is challenging to completely balance the data quantity across different items. Although some observations have low frequency, they still contain valuable information. For instance, even though smaller or remote states may have limited data, their voter preferences are still significant under the Electoral College system. Completely deleting these observations would diminish these states’ representation in the overall analysis, weakening the model’s accuracy in reflecting support rates. Additionally, we aim to cover as many pollsters and states as possible with the model, so we retain even low-frequency observations. Thus, we keep these observations and mitigate the bias they introduce by combining other variables to build a more comprehensive and robust predictive model.

Figure 1 shows the average support rates for the top and bottom ten pollsters, highlighting each pollster’s most commonly used methodology. We can observe that the support rates between the top 5 and bottom 5 pollsters differ by about 10%, which helps us understand the bias introduced by certain pollsters’ influence on support rates, as some pollsters may tend to show higher or lower support rates. This chart also examines the different methodologies used by these pollsters, as varying survey methods (e.g., phone surveys or online panels) may affect the results by reaching different groups of voters. Focusing on these differences aids in assessing the reliability and bias of the data.

Figure 2 compares the average support rates across states with the national average, showing regional differences in voter preferences. Due to the Electoral College system in the U.S. presidential election, variations in state support rates can significantly impact the election outcome. This analysis highlights the differences between state and national polls, helping us understand Trump’s relative support in each state. Paying attention to geographic disparities improves the model’s prediction accuracy.

Figure 3 displays the trend in support rates over time for the top five pollsters with the most observations. We chose these pollsters because a sufficient number of observations is needed to accurately capture trend changes and reduce the impact of extreme values. Each subplot represents a pollster and shows the change in Trump’s support rate since Biden’s withdrawal. This helps us observe the dynamic change in support rates following major events, and we can also see how different pollsters respond to the same event. Analyzing time trends aids in understanding shifts in public opinion, providing more time-sensitive support rate predictions.

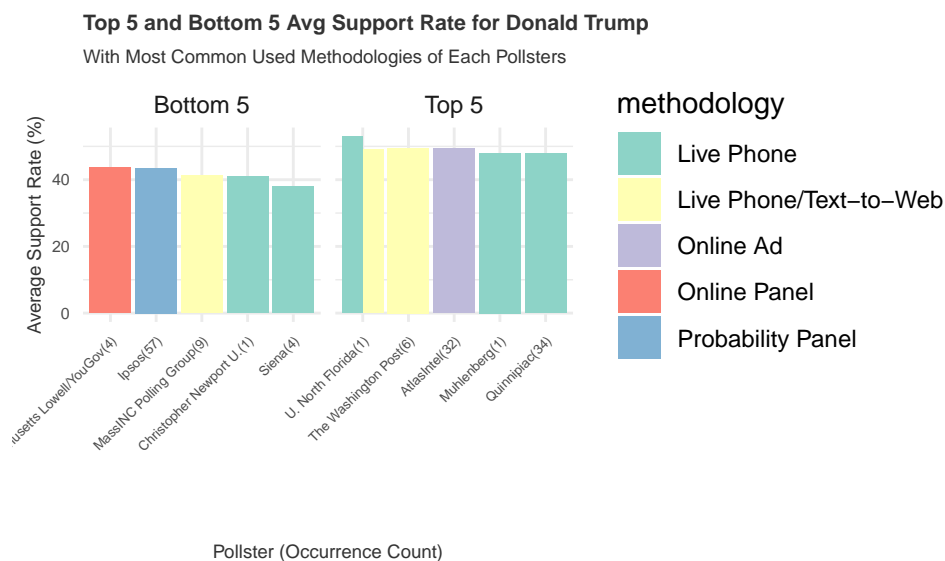


Figure 1: Support Rates for Donald Trump of Ten Pollsters with Their Most Common used Methodology

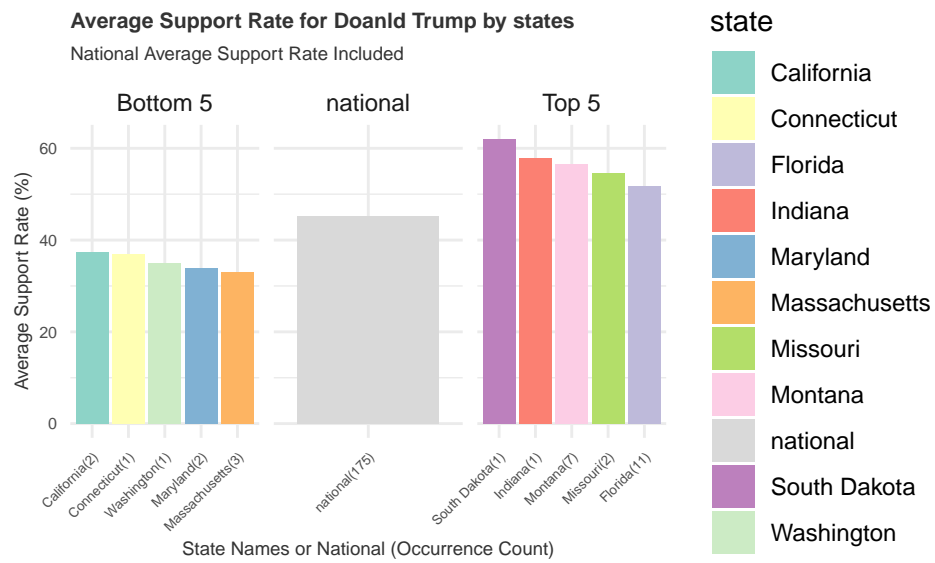


Figure 2: Average Support Rate for Trump of Each Category Based on State

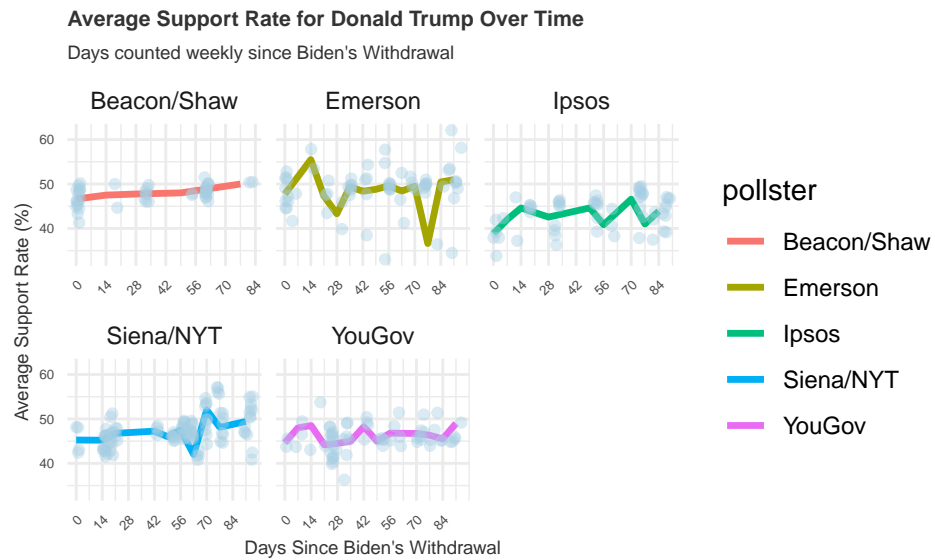


Figure 3: Trend for Average Support Rate for Donald Trump Over Time Since Joe Biden's Withdrawal on 2024 July 21

3 Model

The goal of our modelling strategy is to build a time series model to predict the percentage of support for Trump (pct) over time. This model incorporates multiple variables, including the polling methodology (methodology), the number of days since Biden’s withdrawal from the race on 21 July 2024 (days_since_Biden_Withdrawal), whether the poll is at the national level (is_national), as well as random effects for each state (state) and each pollster (pollster). These variables are chosen to capture the effects of time, polling characteristics, and geographic location on support rates, allowing us to obtain dynamic estimates of support for Trump and to predict the winner of 2024 U.S. presidential election.

Here we briefly describe the Bayesian analysis model used to investigate the factors influencing support for Trump in the polling data. The model incorporates both fixed effects, such as polling methodology and time since Biden’s withdrawal, and random effects to account for variations across states and pollsters. Background details, model assumptions, and diagnostics are included in Appendix B.

3.1 Model set-up

This model is a generalized linear mixed model (GLMM), specifically:

$$y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

Define y_i as the percentage of support (pct) for Trump in poll in the i -th observation, assumed to follow a normal distribution with mean μ_i and variance σ^2 . This model accounts for variations across different methodologies, time since Biden’s withdrawal, and differences between national and state-level polls, with random intercepts for each state and random intercepts and slopes for each pollster.

The linear component of the model is defined as follows:

$$\mu_i = \alpha + \beta_1 \cdot \text{methodology}_i + \beta_2 \cdot \text{days_since_Biden_Withdrawal}_i + \beta_3 \cdot \text{is_national}_i + u_{\text{state}[i]} + (v_{\text{pollster}[i]} + \gamma_{\text{pollster}[i]} \cdot d_i)$$

where:

- y_i is the percentage of support (pct) for Trump in poll i .
- μ_i represents the predicted Trump support for poll i .
- $\alpha \sim \text{Normal}(45, 5)$ is the global intercept for Trump support.
- $\beta_1 \sim \text{Normal}(0, 5)$ represents the effect of polling methodology on Trump support.
- $\beta_2 \sim \text{Normal}(0, 5)$ represents the effect of days since Biden’s withdrawal.
- $\beta_3 \sim \text{Normal}(0, 5)$ distinguishes national polls from state-level polls.

Random effects:

$$u_{\text{state}} \sim \text{Normal}(0, \sigma_{\text{state}})$$

$$v_{\text{pollster}} \sim \text{Normal}(0, \sigma_{\text{pollster}})$$

$$\gamma_{\text{pollster}} \sim \text{Normal}(0, \sigma_{\text{pollster_slope}})$$

where:

- $u_{\text{state}[i]}$ is the random intercept for each state, allowing for state-level variations in Trump support.
- $v_{\text{pollster}[i]}$ is the random intercept for each pollster, capturing pollster-specific effects.
- $\gamma_{\text{pollster}[i]}$ is the random slope associated with the time variable (`days_since_Biden_Withdrawal`) for each pollster.

3.1.1 Model justification

3.1.1.1 Expectation of relationships between variables

Our model applies a hierarchical Bayesian approach, based on the framework proposed by Gelman et al. (2020). This method builds a layered structure that combines polling data collected by different organizations using various methodologies with regional and temporal trends to more accurately predict support rates on Election Day. Running within a Bayesian framework, our model employs a random walk prior to reflect the dynamically changing support rates following Biden’s withdrawal in 2024. This prior distribution is based on historical election trends and the latest polling data, allowing the model to continuously update as new data becomes available.

We anticipate that different polling **methodology** may affect the estimation of support rates. While we cannot precisely determine which polling method might lean toward which demographic group, it is plausible to assume that methods like “online panels” or “phone surveys” might reach different segments of voters, thus introducing some bias into the support rate estimates. For instance, online surveys may engage more frequent internet users, while phone surveys may cover more traditional media users. This bias could lead to varying estimates of support for Trump. By introducing the methodology variable into the model, we aim to correct for these systematic biases associated with different polling methods. Specifically, by controlling for the estimation results across different methods, the model can isolate the biases introduced by each method, allowing the final support rate estimate to better reflect the preferences of the overall voter population. This way, the model can capture potential differences across methods when incorporating data from various polling sources, thereby reducing the systematic error caused by these differences.

For the variable `days_since_Biden_Withdrawal`, we hypothesize that, over time, Trump’s approval ratings may fluctuate due to shifts in the election landscape, policy discussions, and changes in voter sentiment. This hypothesis is based on the dynamic nature of polling data: as

elections near their end, poll results typically converge, reflecting more accurate forecasts for election day (see Blumenthal’s Polls, Forecasts, and Aggregators). The Economist’s forecast model explains this trend, noting that as election day approaches, voters are more likely to solidify their choices, and polling results tend to capture actual public opinion more accurately while the effects of short-term dynamics diminish (How This Works, The Economist). The time factor not only captures shifts in approval ratings as election day draws closer but also allows us to observe how key political events impact approval ratings. For example, candidate debates, policy announcements, or major international events could significantly influence voter preferences at various points in time (see Blumenthal 2010). Thus, we anticipate that the `days_since_Biden-Withdrawal` variable may exhibit a positive or negative trend, depending on the political climate and voter responses to Trump in particular periods.

Through the `is_national` variable, we distinguish between national and state-level polls. We expect that national polls, which typically have larger and more representative samples, are better suited to capturing overall trends, while state-level polls may reflect finer, localized biases and specific regional issues. Because national polls often encompass a broader and more balanced voter base, they provide a more stable forecast of overall voter trends. In contrast, due to geographic and demographic variations, state-level polls often exhibit greater fluctuations in error than national polls. These fluctuations stem from regional issues and diverse demographic characteristics, which may not align fully with national trends (Disentangling Bias and Variance in Election Polls). By introducing `is_national`, this variable enables the model to capture support rate differences arising from geographic and demographic differences. For example, Trump’s support may be higher in traditionally red states, while in swing states, where competition is intense and voter interest is high, his support rates may fluctuate more significantly. This approach helps explain the regional differences in support rates influenced by political leanings and campaign dynamics.

As noted above, differences in state political backgrounds may lead to systematic shifts in support rates; therefore, we included a random intercept (`1 | state`) for each state, with the intention that this random intercept would capture the differing impacts each state has on Trump’s support rate. Similarly, political biases and methodological differences across polling organizations may affect their support rate results. By incorporating (`1 + days_since_Biden-Withdrawal | pollster`), we allow for variability in estimates by different pollsters over time. We anticipate that, due to differences in the stance or methodologies of these organizations, their estimates of Trump’s support rate in the period following Biden’s withdrawal will vary accordingly.

4 Results

4.1 Model Summary

Our results are summarized in a table which is available [here](#). This summary reveals the effects of different polling methods and organizations on our estimates. For instance, there is a small difference in estimates between data collected via online panels and phone surveys. The model adjusts for these differences by using a hierarchical structure, which accounts for inconsistencies caused by methodological and organizational biases. Specifically, variance parameters for different methods (such as online ads vs. phone surveys) indicate the different impacts these methods have on sampling coverage and voter behavior, helping the model balance these differences.

The model's R^2 value is 0.641, indicating that it explains about 64% of the variation in support rates. This shows that the model effectively captures the main patterns in the data, although not every detail. Additionally, the Intraclass Correlation Coefficient (ICC) is 0.9, suggesting a high degree of clustering in the data, meaning that factors like state-specific and polling organization differences significantly contribute to overall variability. Other indicators, such as WAIC (2621.8) and RMSE (2.31), further confirm the model's good fit, showing that it maintains predictive accuracy without overfitting. Also, to balance precision and flexibility, we set a polling error rate of 3% based on historical data. This helps the model capture typical uncertainty without becoming overly confident.

4.2 Prediction Output

From Figure 4, the model predicts that nationwide support for Trump on Election Day, November 5, 2024, will center between 45% to 50% and shows moderate fluctuations. This variability is mainly attributed to national events and differences in polling methods. For instance, nationwide events such as economic shifts or unexpected political occurrences may influence voter sentiment, while differences in polling methods (such as phone surveys versus online panels) may introduce slight fluctuations in estimates due to differences in sample coverage.

At the state level, Figure 5 shows predictions indicate significant variations between states. For example, support in Texas is considerably higher than in California, with a gap of over 10 percentage points, while swing states like Pennsylvania and Michigan display greater volatility. This result further validates the model's robustness; the introduction of a random intercept for each state reflects systematic differences across states. This design aligns with data trends, suggesting that states with diverse geographic and demographic profiles exhibit unique political tendencies.

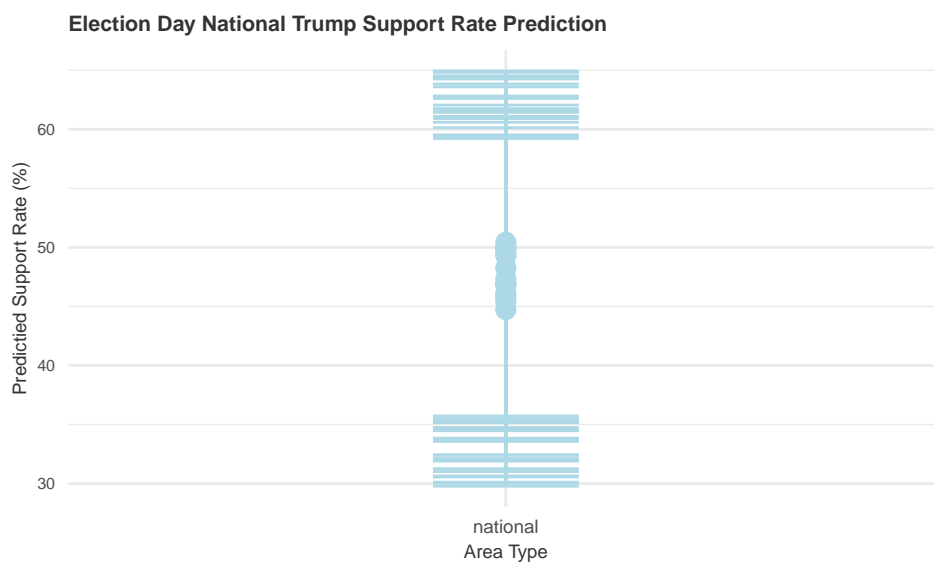


Figure 4: TODO

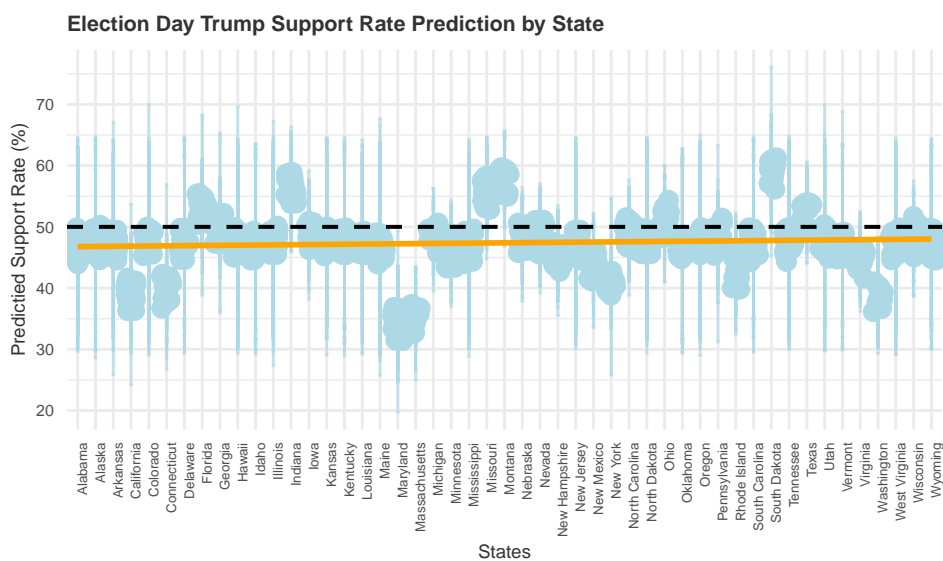


Figure 5: TODO

5 Discussion

In discussing the strengths and limitations of our model, we aim to evaluate its robustness and the challenges it faces in capturing the complexities of voter behavior. Each step, from data collection to model construction, involves certain assumptions and potential biases. Below is a detailed discussion of these aspects:

5.1 Defects of Existing Data

Firstly, a major challenge for the model lies in estimating a large number of parameters with limited polling data. This scarcity of data can make the model highly sensitive to choices in priors and interstate covariance structures. For example, relying on historical patterns from states with limited data may fail to reflect the latest shifts in voter sentiment. This issue is also seen across different polling organizations and survey methods. As Gelman et al. (2020) noted, the model can fluctuate between overconfidence and unreasonable uncertainty. To address this, we could consider incorporating alternative data sources, such as social media sentiment or historical population voting trends, to supplement polling data in underrepresented areas. However, these sources carry their own issues, as they may lack reliability and rigor.

Then, in handling systematic bias introduced by different polling methods, the model accounts for variance by assigning specific variance parameters for each method (e.g., phone interviews vs. online panels). Each method may systematically overestimate or underestimate support for a particular candidate cite . For instance, online panels may attract more politically engaged individuals, leading to biases compared to random phone interviews. To reduce these biases, the model includes variance adjustments specific to each method. Further improvements could involve re-weighting polls based on historical accuracy or introducing dynamic adjustments for newer polling methods. However, as pointed out in 2014 , historical data often include responses to unique events that are not repeatable, making it difficult to assess the accuracy of historical information as a reference. Similarly, the model uses a random walk prior to accommodate fluctuations in public support over time, yet it may struggle to capture abrupt changes driven by political events, such as a debate or controversial statement. The inherent randomness of world events makes it challenging to fully address this problem.

Geographic factors are another focus, with special attention to swing states, which have historically close election results and volatile voter groups. Swing states pose unique challenges for predictive models, as their fluctuations are often inadequately reflected through historical patterns or limited recent polling. Additionally, smaller sample sizes in these states may lead the model to rely too heavily on prior assumptions. Potential solutions include increasing polling frequency in swing states, integrating voter registration data and demographic shifts, or introducing specific polling error adjustments for swing states to enhance model accuracy where small shifts could significantly impact election outcomes.

Furthermore, the polling data we focus on captures voter intent rather than voting probability. Factors such as likelihood of turnout or eligibility are concealed beneath the support rates, and changes in turnout due to major public events may affect election day outcomes. Because these details are not included in the polls, the model cannot fully capture these changes. If the model could dynamically adjust voting behavior estimates based on actual turnout indicators, such as mail ballot requests or early voting numbers, it would better reflect changes in voter behavior and strengthen predictions of the presidential election outcome.

5.2 Polling Deficiencies

5.2.1 Bias and Transparency

The influence of polling agencies' biases and methodological transparency on election predictions has long been a research focus. Every polling agency's choices—whether intentional or unintentional—can introduce biases into results. In highly polarized political environments, factors like question wording, sampling techniques, and demographic weighting can amplify potential biases. For example, research from the Pew Research Center and Stanford University shows that differences in methods, such as educational weighting or nonresponse adjustments, can lead to significant polling biases, especially for politically sensitive figures like Donald Trump (source1).

Another complex issue is “partisan nonresponse bias,” where certain political groups are less likely to participate in polls. This trend has become more prominent in recent years, notably affecting conservative respondents. Pollsters must take extra care in sampling to accurately represent all population segments, which presents particular challenges in practice and makes it hard to eliminate these biases entirely (source2).

Additionally, there is a feedback loop between published poll results and voter behavior: trends shown in polls can influence undecided voters or shift the enthusiasm levels of a candidate's supporters. The presence of “shy” or undecided voters further complicates the challenge of accurate polling. Despite methodological advancements, polling remains prone to errors that are often underestimated compared to the actual unpredictability of elections (source3).

5.2.2 Financial Constraints

Financial limitations are a major barrier to achieving ideal polling coverage, particularly when trying to gather representative polling data across all U.S. states. Polling each state comprehensively is costly, especially in low-population states or states deemed “safe” for one candidate. In these cases, data scarcity means the model largely relies on data from more populous or swing states, which may not fully represent the diversity of the nationwide electorate, introducing potential biases.

According to a Pew Research Center report, polling biases often stem from financial constraints and methodological choices, such as the use of cost-effective online panels, which may not effectively capture perspectives from older, rural, or technologically less-adept groups. These groups often require different recruitment strategies and data collection methods, increasing polling costs (Pew Research Center, 2024) 387†source . Another Pew report highlights that participation in extreme partisan regions is often low, which makes balanced representation more challenging and increases the financial and logistical burden of implementing corrective weighting and sampling techniques 388†source .

In theory, expanding polling coverage could improve model representativeness and accuracy. However, due to the high cost of polling and the statistical law of diminishing marginal returns, this expansion is often constrained by budget and logistical support. The concept of diminishing marginal returns means that as more resources (like time, money, or personnel) are invested in an activity, the additional benefit from each extra unit of investment gradually decreases. For instance, the first slice of pizza might be very satisfying, but by the third or fourth slice, enjoyment decreases. For polling, investing resources to cover more regions initially increases data representativeness significantly, but as major groups are covered, further investments yield smaller improvements, inevitably leaving certain groups underrepresented in the results.

5.3 The Influence of Reality

5.3.1 The Balance Between Model Complexity and Interpretability

While our model integrates multiple parameters and hierarchical structures to address the variability across states and polling methodologies for a more accurate election support forecast, this complexity also introduces significant challenges in interpretability. For non-expert users, understanding how random intercepts or variance parameters influence the final prediction requires a certain level of statistical knowledge. Even though we attempt to make the model's main drivers (e.g., state-level effects or specific methodological impacts) more understandable through simplified summaries and visualizations, several issues remain.

Visualizations often struggle to fully capture the complex structure of our model. The model includes data from 50 U.S. states, multiple polling organizations, and various survey methodologies. If we attempted to display every detail within a chart, the visualizations would become overwhelmingly complex, making it difficult for readers to extract key insights. Conversely, if we limit the display to a subset, such as only highlighting key states or selected polling methods, this approach risks losing the completeness of information. These simplified visualizations fail to convey all sources of variation within the model, potentially leading readers to overlook significant factors and misunderstand the impact of omitted influences. Additionally, this partial representation can foster a feedback loop where oversimplified summaries influence potential voters' perceptions, impacting the trends in unpredictable ways.

5.3.2 The Randomness of the Real World

Voter preferences can shift rapidly due to events like debates, economic changes, or international incidents, challenging election predictions. Regular polling may fail to capture these shifts if not updated frequently, leading to outdated and potentially misleading forecasts. The Harvard Data Science Review highlighted this uncertainty in the 2020 election, where state-level opinion volatility was influenced by economic and educational factors, necessitating models that incorporate these dynamic interactions[HARVARD DATA SCIENCE REVIEW]

With elections occurring only every four years, opportunities to validate prediction models are limited. Standard techniques like cross-validation are challenging to implement due to data scarcity, pushing models to rely heavily on historical data, which risks overfitting to past trends. Harvard Data Science Review notes that this reliance can result in models exhibiting excessive certainty or unreasonable uncertainty, especially when recent polling data is lacking

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

B.2 Diagnostics

References

- Alexander, Rohan. 2024. “Starter Folder.” https://github.com/RohanAlexander/starter_folder.git.
- FiveThirtyEight. 2024. “2024 Presidential General Election Polls.” <https://projects.fivethirtyeight.com/polls/president-general/2024/national/>.
- Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. “rstanarm: Bayesian applied regression modeling via Stan.” <https://mc-stan.org/rstanarm/>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. <https://broom.tidymodels.org/>.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemond, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. <https://readxl.tidyverse.org>.
- Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. <https://dplyr.tidyverse.org>.
- Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://cran.r-project.org/web/packages/knitr/index.html>.