Analyzing Trends in Trump Support for the 2024 U.S. Presidential Election*

Stable Trump Support with Regional Variations: Insights from Interaction Modeling on FiveThirtyEight Polling Data

Xuanang Ren Caichen Sun

November 4, 2024

The 2024 U.S. Presidential election features a close race between Vice President Kamala Harris and former President Donald Trump. In this study, we analyzed public support for Trump using data from FiveThirtyEight, employing an interaction model to assess trends across key battleground states. Our model incorporates time, state, and poll quality as predictors to capture both regional and methodological influences on support levels. The findings reveal a generally stable support base for Trump, with minor fluctuations across states like Michigan and Georgia. This stability in support, despite methodological variations, highlights the resilience of Trump's voter base and underscores the importance of regional and pollster adjustments in electoral forecasting.

1 Introduction

The 2024 U.S. Presidential election sees Donald Trump once again vying for the White House amid a deeply polarized political climate. Public support for Trump, particularly in key battleground states, remains a critical focus for analysts and strategists alike. This paper explores trends in Trump's support using data compiled by FiveThirtyEight, focusing on variations in polling data across states and over time. By utilizing an interaction model, we analyze how factors like state, time, and poll quality shape Trump's popularity leading up to the election.

Our analysis examines how Trump's support changes over time within key states, as well as how different polling methodologies and poll qualities influence reported support levels. We focus on battleground states such as Arizona, Georgia, Michigan, North Carolina, Pennsylvania, and

*Code and data are available at: https://github.com/JanosRenAI/2024-US-Presidential-Election.git.

Wisconsin, as these states hold significant sway in determining the election outcome. The true underlying effect of factors such as time, state, and poll quality on Trump's support remains unknown, but our model provides an informed approximation of these impacts across regions and methodologies.

Our findings suggest that Trump's support remains relatively stable across polls, with only slight increases or fluctuations in certain states. For example, Michigan and Georgia demonstrate a modest upward trend in Trump's support, while North Carolina and Wisconsin show more variability. Despite the differences observed in individual states, the overall trend indicates a resilient base of support for Trump, relatively unaffected by polling variations or differences in poll quality.

In a time of intense political division, understanding public sentiment and trends in candidate support has become more essential than ever for accurate electoral forecasting. Our analysis provides a nuanced perspective on Trump's support, emphasizing the importance of adjusting for regional dynamics and poll quality in interpreting polling data. This study contributes to a clearer picture of voter sentiment and enhances the reliability of predictions for the upcoming election.

The remainder of this paper is organized as follows: Section 2 describes the data and cleaning process, including visualizations and summary statistics. Section 3 outlines the interaction model used to assess Trump's support across different states and time periods. Section 4 presents the results, including tables and visualizations of national and state-level trends. Section 5 discusses the implications of these findings and addresses limitations in our approach. The appendix contains additional details on an idealized methodology for survey design and poll aggregation, enhancing future electoral forecasting.

2 Data

2.1 Overview

This study employs various R packages (R Core Team 2023) for data cleaning and analysis, including libraries from tidyverse (**tidyverse?**), ggplot2 (**ggplot2?**), palmerpenguins (Horst, Hill, and Gorman 2020), dplyr (**dplyr?**), readr (**readr?**), testthat (**testthat?**), rstanarm (Goodrich et al. 2022), janitor (**janitor?**), and modelsummary (**modelsummary?**). Additionally, the arrow package (**arrow?**) is used for efficient data handling and storage.

The dataset used in this analysis is sourced from FiveThirtyEight's publicly available "Presidential general election polls (current cycle)," which compiles polling data on the 2024 U.S. presidential election. This dataset offers a comprehensive view of national polling trends, including pollster information, sample sizes, candidate preferences, and polling methodologies.

By aggregating data from a wide range of reputable pollsters, FiveThirtyEight applies a rigorous methodology to clean and standardize the dataset, enhancing its reliability for statistical modeling.

In terms of broader context, this dataset is situated within the field of electoral forecasting, where accurate polling data is essential for making informed predictions. It is specifically valuable as it reflects the most current cycle, allowing the analysis to capture evolving voter sentiments as the election date approaches. While similar datasets from past election cycles exist, such as historical polls from previous presidential elections or datasets from other polling aggregators, they were not chosen for this analysis due to the recency and specificity of the FiveThirtyEight data. Utilizing outdated or less comprehensive datasets could introduce biases that do not accurately represent current voter dynamics. Consequently, this dataset provides the most relevant snapshot of the electorate's preferences, making it ideal for forecasting purposes in this context.

2.2 Clean data

The final cleaned dataset includes variables essential for analyzing Donald Trump's support in the 2024 U.S. presidential election polls. Each row in this dataset represents a single poll observation, capturing information about the pollster, quality score, sample size, state (or region), end date of the poll, and the percentage of respondents indicating support for Trump.

In our analysis, we initiated the data cleaning by importing the raw dataset, which consists of various poll records. The cleaning steps involved standardizing column names for consistency using the clean_names() function from the janitor package. We then focused on data related to Donald Trump, applying a filter to include only records where Trump was a candidate and the numeric grade of the poll was above the established threshold of 2.7. To ensure the integrity of the dataset, we managed missing values by assigning 'National' to polls lacking state information and grouped state-based records, reclassifying those with fewer than 60 polls under a collective category labeled 'Other'. This step helps in maintaining a robust sample size for meaningful analysis.Further, we converted date formats to ensure consistency and filtered the data to include only records after Trump's official candidacy declaration. Outliers in percentage values, specifically those outside the 0-100% range, were removed to maintain data accuracy. Lastly, we calculated the number of Trump supporters derived from the percentage and sample size, retaining only data from pollsters with more than 20 polls for reliability. The cleaned data was then saved in CSV and Parquet formats, facilitating subsequent analysis and visualization stages.

2.3 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

Table 2: Summary Statistics of Outcome Variable

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
pct	66	0	48.0	3.3	33.3	48.0	57.9
num_trump	147	0	443.1	133.3	266.0	432.0	1022.0

2.4 Outcome variables

2.4.1 Example

The table below Table 1 is an example of each dependent variable

Table 1: Outcome Variables

Outcome Variable	Example
Percentage of Votes (pct)	49.2
Actual Supporter Count of Trump (num_trump)	708

2.4.2 Summary Statistics

The summary statistics Table 2 indicate that the pct values vary from a minimum of 32.0% to a maximum of 57.9%, with an average (Mean) support rate of 46.6% and a standard deviation (SD) of 3.8. This suggests a relatively consistent level of support for Trump across the various polls included in the study. According to the summary statistics, the number of Trump supporters estimated by polls ranged from 161 to 2123, with an average of 528 supporters per poll and a standard deviation of 277.8. The median of 471 indicates that half of the polls estimated fewer than 471 Trump supporters, highlighting variability in poll sizes and support levels. This distribution underscores the need for this transformation, providing a more intuitive understanding of support levels when comparing polls with different sample sizes.

2.4.3 Percentage of Votes ('pct')

The percentage of respondents who supported Trump in each poll. To ensure validity, outliers with pct values outside the range of 0-100 were removed. This variable represents the core measure of Trump support used in modeling.

2.4.3.1 Percentage distribution

```
NULL
```

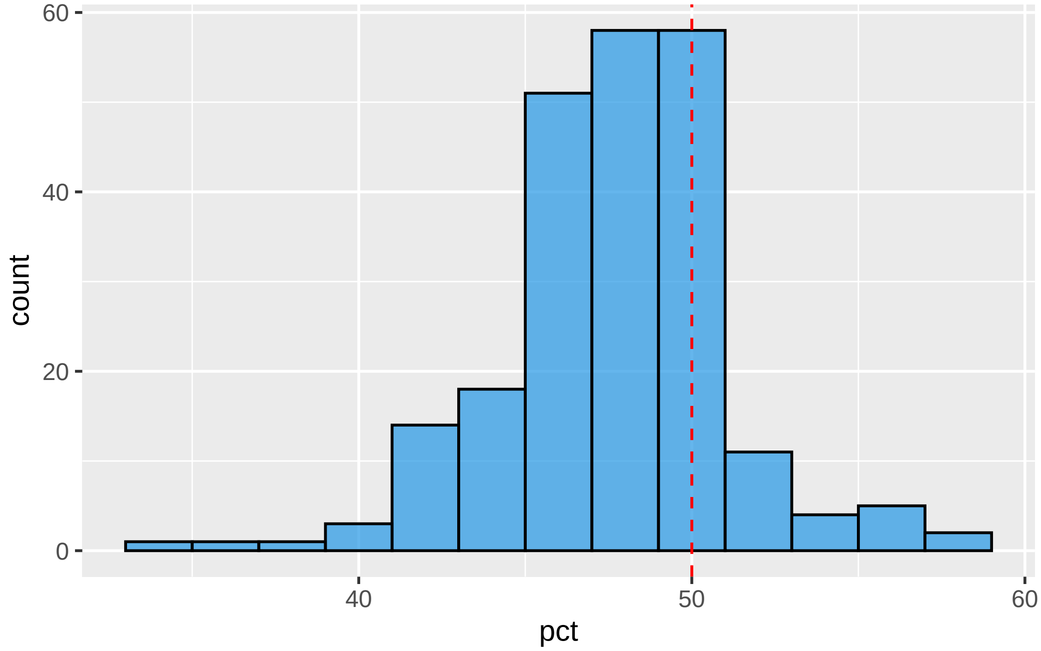

Figure 1: Percentage distribution of support for Donald Trump across various poll responses. The red dashed line at the 50% mark serves as a reference, indicating the threshold for majority support.

The primary variable examined in this study is the percentage of public support for Donald Trump. As illustrated in Figure Figure 1, the x-axis represents support percentages, spanning from around 20% to 70%, while the y-axis shows the frequency of poll responses at each support level. A red dashed line at the 50% mark serves as a reference point, highlighting the threshold for achieving majority support.

The distribution indicates that most poll results cluster around a 48% support level, reflecting substantial backing from the electorate. Fewer polls report support exceeding 55%, suggesting that, while Trump has a dedicated base, a significant portion of voters remain undecided or opposed. This visualization emphasizes both the concentration and range of support levels across the surveyed population, providing insights into the variability of public opinion.

2.4.3.2 Actual Supporter Count of Trump ('num_trump')

A constructed variable that translates pct into actual supporter counts, calculated as (pct / 100) * sample_size. This transformation is crucial for accurately estimating Trump support in terms of respondent numbers, which is often more intuitive for interpreting support levels across polls of varying sizes.

In the histogram Figure 2, the x-axis is labeled "Number of Trump Votes," representing the transformed supporter counts, and the y-axis is labeled "Frequency," indicating the number of occurrences for each vote range within the dataset. The majority of the polls (as shown by the height of the bars) indicate fewer than 1000 Trump votes, demonstrating a heavy left-skew in the distribution. A few polls exhibit considerably higher Trump support, extending towards the right end of the plot, but these are less frequent.

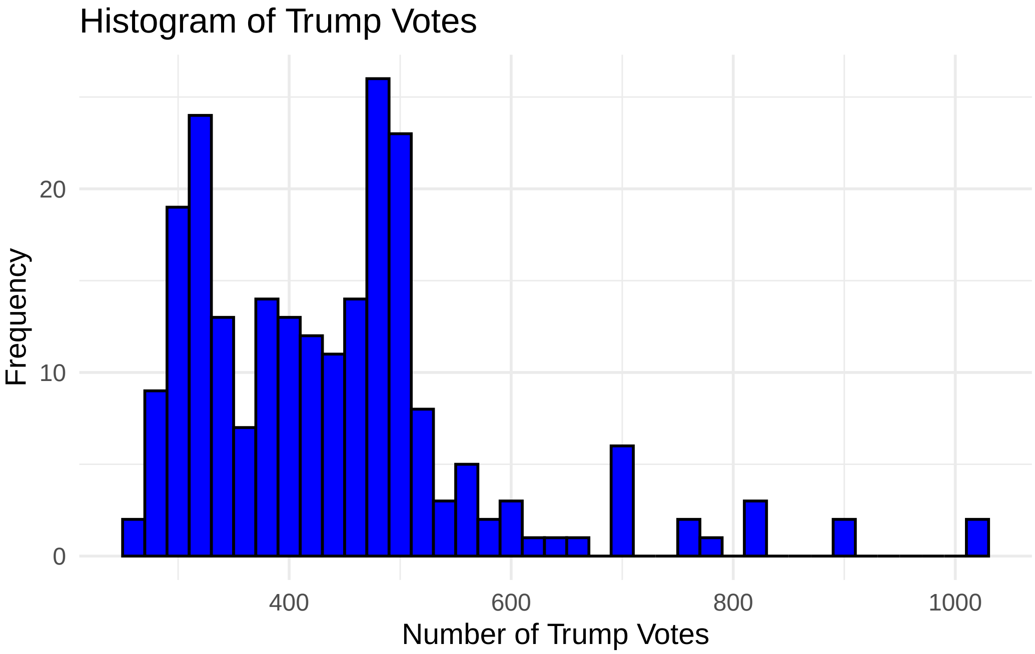

Figure 2: Distribution of Actual Supporter Count of Trump

2.5 Predictor Variables

2.5.1 Example

The table below Table 3 is an example of each dependent variable

Table 4: Summary Statistics of Predictor Variable

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
numeric_grade	4	0	2.9	0.1	2.7	2.9	3.0
pollscore	4	0	-1.1	0.3	-1.5	-1.1	-0.5
sample_size	76	0	921.9	262.6	617.0	905.0	2048.0

Table 3: Predictor Variables

Predictor Variable	Example
Pollster (pollster)	AtlasIntel
Poll Quality (numeric_grade)	2.7
Pollscore (pollscore)	-0.8
State (state)	Arizona
End Date (end_date)	2024/10/17
Sample Size (sample_size)	1440

2.5.2 Summary Statistics

2.5.3 pollster ('pollster')

This categorical variable identifies the organization conducting each poll. Only pollsters with a minimum of 20 polls are included, ensuring consistency and reliability in polling patterns across the dataset. By limiting the dataset to higher-frequency pollsters, we enhance its representativeness and statistical validity.

2.5.4 Poll Quality ('numeric_grade')

This variable is a broad quality measure assigned to each pollster by FiveThirtyEight, reflecting the historical accuracy and reliability of the pollster's work. A higher numeric_grade indicates a pollster with a stronger track record, and the threshold of 2.7 was set in this analysis to include only those pollsters generally recognized as reliable. This variable is therefore used as a filter to ensure that only higher-quality pollsters contribute to the dataset.

2.5.5 Pollscore (‘pollscore’)

In contrast, pollscore is a specific, individual score for each poll within the dataset. It accounts for factors unique to that particular poll, such as timing, methodology, and sample structure. The pollscore allows the model to distinguish between the perceived quality of different polls conducted by the same pollster. Even a pollster with a high numeric_grade might produce a poll with a lower pollscore due to circumstances like sample size limitations or the recency of data collection.

2.5.6 State (‘state’)

The geographic focus of the poll, which defaults to “National” for national-level polls or “Other” for states with limited polling data. This grouping strategy reduces fragmentation from states with sparse data while retaining the robustness of state-level polling insights.

2.5.7 End Date (‘end_date’)

The date when polling data collection ended, formatted for time-series analysis. Only polls after Trump’s formal declaration (July 21, 2024) are included, narrowing the dataset to reflect recent sentiment aligned with active campaign periods.

2.5.8 Sample Size (‘sample_size’)

The number of respondents in each poll, which provides a basis for understanding the scale of each poll’s outreach. This variable contextualizes the precision of pct measurements, as larger sample sizes generally yield more reliable estimates.

2.5.8.1 Sample Size Distribution

The sample size in each poll represents the number of respondents surveyed, serving as a key indicator of the data’s reliability and representativeness. Figure Figure 3 illustrates the distribution of sample sizes across various polls. The histogram shows a right-skewed distribution, where the majority of polls feature smaller sample sizes, and the frequency diminishes as the sample size increases. This right skew suggests that while larger sample sizes are present, they are less common in the dataset.

The distribution peaks at approximately 1,000 respondents, indicating this as the most frequent sample size among the polls analyzed. This peak highlights a standard practice in polling methodology, where sample sizes around 1,000 are generally deemed sufficient to provide statistically meaningful insights. Consequently, the sample sizes across the dataset appear robust enough to yield reliable information, despite variability. This distribution provides confidence in the dataset's representativeness, given that even the smaller samples are generally adequate for capturing broad public opinion trends.

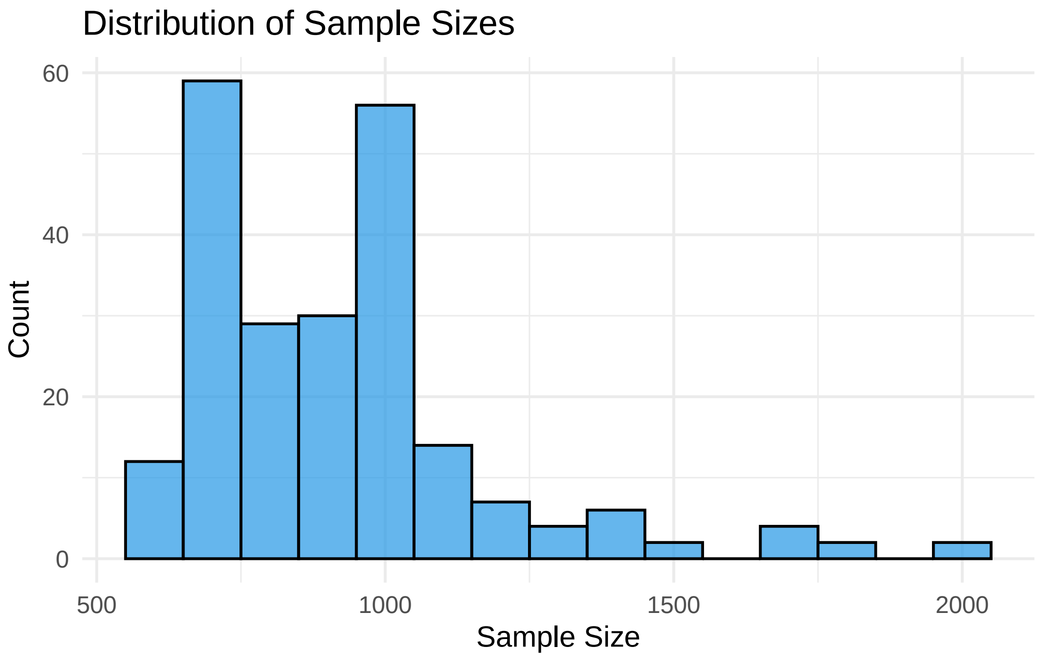

Figure 3: Distribution of Sample Sizes Across Various Polls

2.6 Trends in Trump Support

This study examines trends in public support for Donald Trump over time, as reported by various polling organizations. By tracking shifts in support across different time points and pollsters, the analysis offers insights into the dynamics of public opinion throughout the election cycle. The observed fluctuations in Trump's support likely correlate with major campaign events, policy changes, or shifts in voter sentiment.

Additionally, discrepancies between polling organizations highlight methodological differences, such as sampling techniques and demographic weighting, which may affect support estimates. These variations underscore the importance of considering pollster methodology and potential biases when interpreting polling data. This framework thus provides a nuanced perspective on tracking candidate support, contributing to a more accurate and comprehensive understanding of election trends.

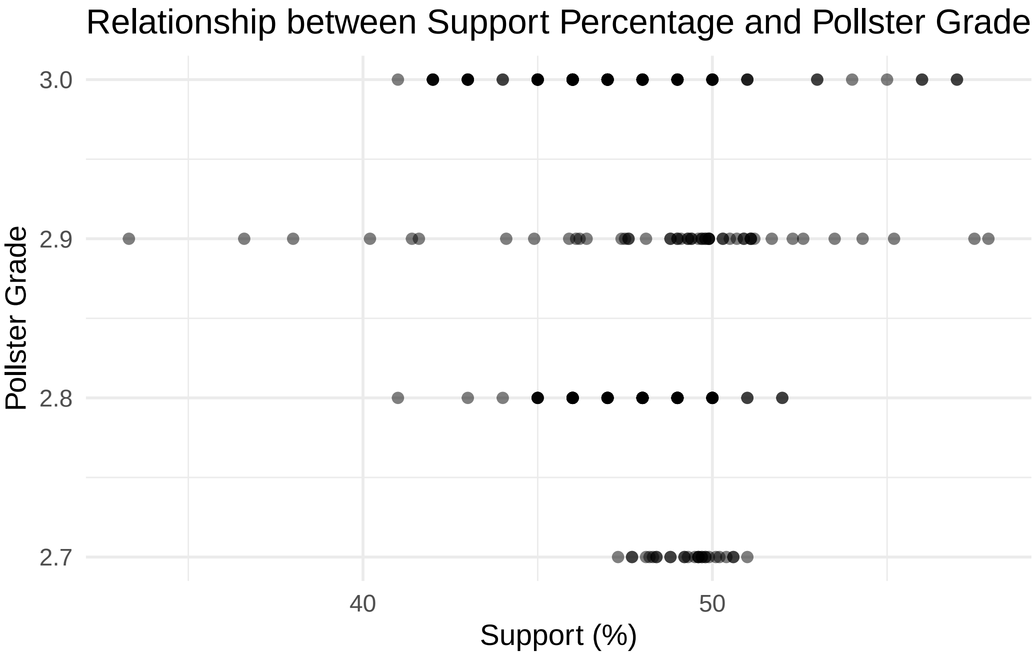

Figure 4: Distribution of Correlation Between Support and Pollster Grade Across Polls

2.7 Trend of Trump's support over time by pollster

The chart Figure 5 shows the trend in support for Donald Trump as estimated by different polling agencies from August to October, with each agency represented by a unique color. The y-axis shows support percentages, and the x-axis represents dates. Overall, Trump's support remains relatively stable, fluctuating within a narrow range of about 40% to 60%, indicating a steady support base with no significant changes over time.

There are minor differences between polling agencies. For example, Quinnipiac and Siena/NYT display more noticeable fluctuations, possibly due to differences in methodology or sampling. In contrast, AtlasIntel shows a steadier trend, suggesting more consistent sampling or modeling practices. These variations highlight the impact of methodology on polling results, with factors like sample demographics, weighting, and polling frequency potentially contributing to the differences between agencies.

Overall, the chart suggests that while individual polls may vary slightly, Trump's support level remains stable throughout the period. This stability implies a resilient support base, though the small differences across agencies underscore the importance of considering multiple polls together for a more accurate view of public opinion trends.

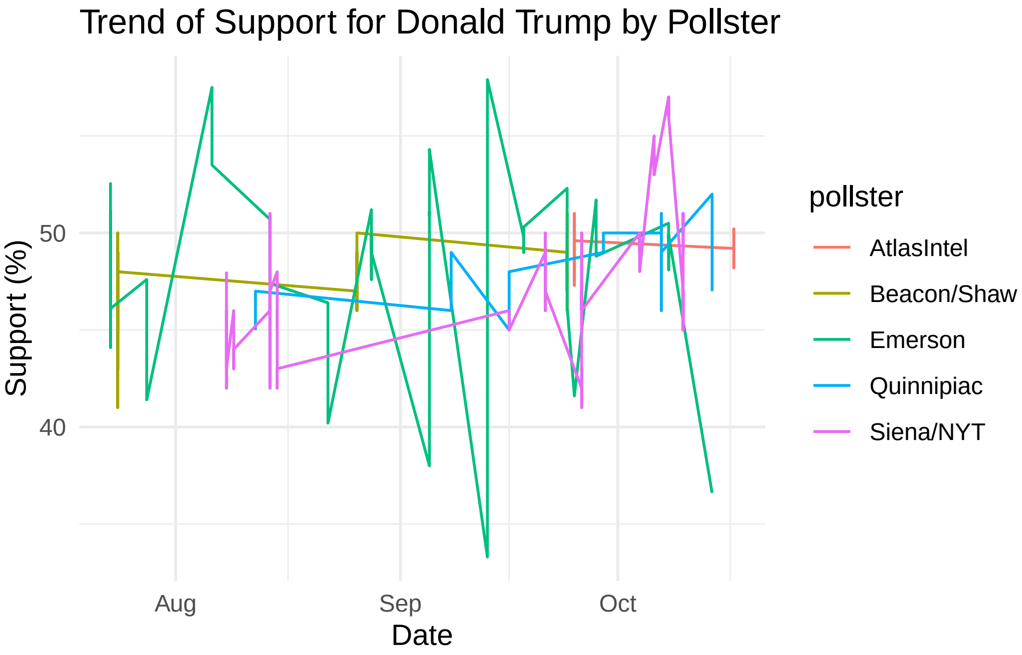

Figure 5: Distribution of Correlation Between Support and Pollster Grade Across Polls

3 Model

In this part of our study, we focus on overcoming the biases and disparities in polling data to develop a reliable predictive model. The primary challenge involves choosing a model that strikes the right balance between complexity and accuracy, making sure it effectively reflects the intricacies of polling data without succumbing to overfitting. For this purpose, we have meticulously analyzed various model configurations to select the one best suited for our forecasting needs.

In this section, we focus on refining our approach to managing polling data's inherent biases and variances to build a dependable predictive model. Selecting a model involves achieving a delicate balance between sufficient complexity to capture essential data features and simplicity to prevent overfitting. Our extensive analysis of various models has led us to adopt the most suitable configuration for forecasting.

The primary objective of our model is to accurately predict the percentage of support for a political candidate, in this case, Donald Trump, using historical polling data. This involves addressing data discrepancies and incorporating relevant features that influence polling outcomes.

In this part of our study, we focus on refining our approach to handling the complexities of polling data across different U.S. states to predict support for a political candidate, Donald

Trump. Our primary challenge is to construct a model that captures state-specific trends in support over time while remaining interpretable and manageable in complexity. We carefully considered several model configurations to choose a balanced approach that aligns with the polling data's nuances.

The goal of our modeling strategy is twofold. Firstly, it aims to accurately reflect how Trump's support varies over time within each state. Secondly, it seeks to adjust for polling quality by incorporating poll scores, which help us control for variations in methodology across different polls. This approach ensures that the model is not overly simplistic, as it includes interaction terms that capture state-level trends, yet remains focused on core features relevant to the data.

In this section, we present **Model 5**, an interaction model that includes both time (`end_date`) and state (`state`) as interacting predictors, as well as poll scores (`pollscore`) to account for differences in poll quality. We chose this model after extensive analysis due to its ability to reveal nuanced, state-specific trends in Trump's support over time.

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in **?@sec-model-details**.

3.1 Model set-up

Let y_i represent the percentage of Trump support in poll i, measured as the proportion of respondents indicating support. The models are defined as follows:

3.1.1 Final Model (Model 5): State Interaction Model

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 \times \text{end\_date}_i + \beta_2 \times \text{state}_i + \beta_3 \times \text{end\_date}_i \times \text{state}_i$$
$$\alpha \sim \text{Normal}(0, 2.5), \quad \beta \sim \text{Normal}(0, 2.5), \quad \sigma \sim \text{Exponential}(1)$$

- y_i represents the percentage of Trump support in poll i, expressed as the proportion of respondents indicating support in that poll.
- μ_i represents the expected percentage of Trump support in poll i, as predicted by the model.
- α is the intercept, representing the baseline level of Trump support when other predictors are at zero.
- β_1 is the coefficient for `end_date`, capturing the effect of the poll's recency on Trump support, with later polls potentially indicating shifts in support over time.
- β_2 is the coefficient for `state`, capturing the effect of being in a specific state on Trump support, accounting for state-level variations.

- β_3 is the coefficient for the interaction term `end_date` × `state`, capturing how the relationship between recency of the poll and Trump support may vary across states, allowing for different trends by state.
- σ is the standard deviation of the model's residuals, representing the variability in Trump support that is not explained by the model.
- ϵ_i denotes the error term for poll i, assumed to follow a normal distribution with mean 0, capturing random variation in Trump support that is not explained by the included predictors.

This model includes an interaction term between state and date to account for how support trends might differ significantly across states.

3.1.2 Model 1: Simple Linear Model

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 \times \text{end\_date}_i$$
$$\alpha \sim \text{Normal}(0, 2.5), \quad \beta_1 \sim \text{Normal}(0, 2.5), \quad \sigma \sim \text{Exponential}(1)$$

- y_i represents the percentage of Trump support in poll i, as a proportion of respondents indicating support.
- μ_i represents the expected percentage of Trump support in poll i, as predicted by the model.
- α is the intercept, representing the baseline level of Trump support.
- β_1 is the coefficient for `end_date`, capturing the effect of time (recency of the poll) on Trump support.
- σ represents the residual standard deviation, accounting for variability not explained by the model.
- ϵ_i denotes the error term, assumed to follow a normal distribution with mean 0.

This model estimates the overall trend of Trump support over time without considering the effects of other variables.

3.1.3 Model 2: Pollster Effect Model

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 \times \text{end\_date}_i + \sum_{j=1}^{J} \beta_{2j} \times \text{pollster}_{ij}$$
$$\alpha \sim \text{Normal}(0, 2.5), \quad \beta \sim \text{Normal}(0, 2.5), \quad \sigma \sim \text{Exponential}(1)$$

- y_i represents the percentage of Trump support in poll i, measured as the proportion of respondents indicating support.
- μ_i represents the expected percentage of Trump support in poll i, as predicted by the model.
- α is the intercept, representing the baseline level of Trump support.
- β_1 is the coefficient for `end_date`, capturing the effect of recency on Trump support.
- β_{2j} are the coefficients for `pollster`, with each β_{2j} capturing the effect of a specific pollster on Trump support, accounting for variations in methods and biases among different pollsters.
- σ represents the residual standard deviation of the model.
- ϵ_i denotes the error term, assumed to follow a normal distribution with mean 0.

This model introduces a variable for pollster to account for variations in polling methods and biases among different pollsters.

3.1.4 Model 3: Sample Size and Poll Score Model

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 \times \text{end\_date}_i + \beta_2 \times \text{sample\_size}_i + \beta_3 \times \text{pollscore}_i$$
$$\alpha \sim \text{Normal}(0, 2.5), \quad \beta \sim \text{Normal}(0, 2.5), \quad \sigma \sim \text{Exponential}(1)$$

- y_i represents the percentage of Trump support in poll i, as a proportion of respondents indicating support.
- μ_i represents the expected percentage of Trump support in poll i, as predicted by the model.
- α is the intercept, representing the baseline level of Trump support.
- β_1 is the coefficient for `end_date`, capturing the effect of time on Trump support.
- β_2 is the coefficient for `sample_size`, representing the effect of sample size on Trump support, which may adjust the estimate's reliability.
- β_3 is the coefficient for `pollscore`, capturing the effect of the poll's quality or credibility score on Trump support.
- σ represents the residual standard deviation of the model.
- ϵ_i denotes the error term, assumed to follow a normal distribution with mean 0.

This model considers additional variables such as the size of the sample and the quality score of the poll, providing more granularity.

3.1.5 Model 4: Hierarchical Pollster Model

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 \times \text{end\_date}_i + \beta_2 \times \text{pollster}_i$$

$$\alpha \sim \text{Normal}(0, 2.5), \quad \beta_1, \beta_2 \sim \text{Normal}(0, 2.5) \text{ (with pollster-specific adjustments)}, \quad \sigma \sim \text{Exponential}(1)$$

- y_i represents the percentage of Trump support in poll i, as a proportion of respondents indicating support.
- μ_i represents the expected percentage of Trump support in poll i, as predicted by the model.
- α is the intercept, representing the baseline level of Trump support.
- β_1 is the coefficient for `end_date`, capturing the effect of time on Trump support.
- β_2 is the coefficient for `pollster`, representing pollster-specific effects, which allow for variability in Trump support across different pollsters.
- σ represents the residual standard deviation of the model.
- ϵ_i denotes the error term, assumed to follow a normal distribution with mean 0.

In this hierarchical model, pollster effects are modeled not just as fixed effects but are considered to vary across different polls, introducing random effects for each pollster.

We run these models in R using the `rstanarm` package (Goodrich et al. 2022). We use the default priors from `rstanarm` for our Bayesian analyses.

3.1.6 Model justification

3.1.6.1 Overview

In statistical analysis of polling data, selecting an appropriate model is crucial for valid inferences. Our analysis aimed to capture trends in support for Trump, considering temporal changes and regional variations. Given the complexity of electoral dynamics and polling methodologies, our model needed to balance simplicity for interpretability and sufficient complexity to capture essential features in the data.

3.1.6.2 Model Complexity and Appropriateness

We began with simpler models that set a baseline for understanding temporal trends and pollster effects. However, these models, including linear adjustments for time and pollster (Model 1 and Model 2), were inadequate for our dataset, which contains significant regional disparities and interaction effects between state and time. Model 3 introduced sample size and poll score, improving model fit slightly but still not addressing state-based variations.

3.1.6.3 Advantages of Model 5: State Interaction Model

Model 5 incorporates interactions between states and polling dates, significantly improving our understanding of the data. This model is defined by the equation:

$$y_i | \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_1 \times \text{end\_date}_i + \beta_2 \times \text{state}_i + \beta_3 \times \text{end\_date}_i \times \text{state}_i$$

$$\alpha \sim \text{Normal}(0, 2.5), \quad \beta \sim \text{Normal}(0, 2.5), \quad \sigma \sim \text{Exponential}(1)$$

This model is superior because it allows each state's support trend to have its own trajectory over time, crucial for accurately modeling political behaviors in diverse geographical regions. This approach is particularly valuable in predicting outcomes in swing states, where small shifts can have significant electoral implications.

3.1.6.4 Justification of Complexity

The complexity added by the interaction terms is justified by their significant impact on model performance. R-squared values indicate that Model 5 explains more variability than simpler models, and although the AIC is slightly higher, the improvement in model fit is substantial. This trade-off between complexity and fit is a common challenge in statistical modeling, and in our case, the additional complexity is warranted given the analytical benefits.

3.1.6.5 Data Considerations

Our dataset comprises multiple polls conducted at different times across various states. Each state has unique political dynamics that could influence polling results, making the interaction between state and time critical. The decision to include state interactions is supported by exploratory data analysis showing significant variance in Trump's support across states over time.

3.1.6.6 Bayesian Priors and Sensitivity

The priors for the coefficients were set as Normal(0, 2.5), reflecting a neutral starting point and allowing the data to inform the posterior distributions significantly. This choice of priors is conservative, avoiding overconfident assumptions in a highly uncertain electoral context. Sensitivity analyses were conducted to ensure that the results are robust across different prior specifications, confirming that the model's conclusions are not unduly influenced by the choice of priors.

3.1.6.7 Model Implementation and Validation

Implemented in R using the `rstanarm` package, Model 5's Bayesian framework allows for rigorous uncertainty quantification and model checking. Convergence diagnostics were carefully monitored using trace plots and the Gelman-Rubin statistic, ensuring that the sampling adequately explored the posterior distribution. Out-of-sample validation involved partitioning the data into training and testing sets, where Model 5 consistently showed lower RMSE and better predictive accuracy compared to its simpler counterparts.

3.1.6.8 Alternative Models and Final Choice

While we considered other models, including those focusing on individual effects of pollsters or samples, Model 5 was ultimately selected due to its comprehensive approach and superior performance. Alternative models were either too simplistic to capture the crucial state-time dynamics or did not substantially improve upon the predictive power of Model 5.

3.1.7 Conclusion

Model 5 represents the best balance between complexity and explanatory power for our study on Trump support trends. Its ability to effectively model the interaction between state dynamics and temporal trends, supported by robust statistical validation, makes it the most suitable choice for our analysis, providing insightful and reliable predictions.

4 Results

Our model predicts that support for Donald Trump in the 2024 U.S. presidential election remains within a narrow range across various polls. As shown in Figure 1, Trump's support has averaged around 48%, with the highest poll results slightly exceeding 55%. The clustering of polls near the 48% mark highlights substantial backing among likely voters. However, relatively few polls have recorded support levels above 55%, suggesting that while Trump has a strong base, a significant portion of voters remains undecided or opposes him.

Figure Figure 6 presents the trend of Donald Trump's support percentage over time in various states, including Arizona, Georgia, Michigan, North Carolina, Pennsylvania, Wisconsin, and an aggregated "Other" category. Each sub-plot represents a different state, with gray dots marking individual poll observations and colored trend lines showing the smoothed support trend for each state.

The data reveals several state-specific patterns in Trump's support. For example, Michigan and Georgia show a noticeable upward trend, indicating growing support over time. In contrast, states like Arizona and North Carolina exhibit more stable trends, with only slight increases in support. The trend lines for Pennsylvania and Wisconsin indicate variability but generally suggest a steady support level around 47-50%.

The distinct color for each state allows for easy comparison of Trump's support trajectories across regions. The "Other" category, which aggregates less frequently polled states, shows a stable trend with minimal fluctuations.

This figure underscores the importance of analyzing Trump's support on a state-by-state basis, as regional differences provide insights into the dynamics of his base. Overall, while there are slight increases in some states, Trump's support appears stable across most regions as the election period progresses.

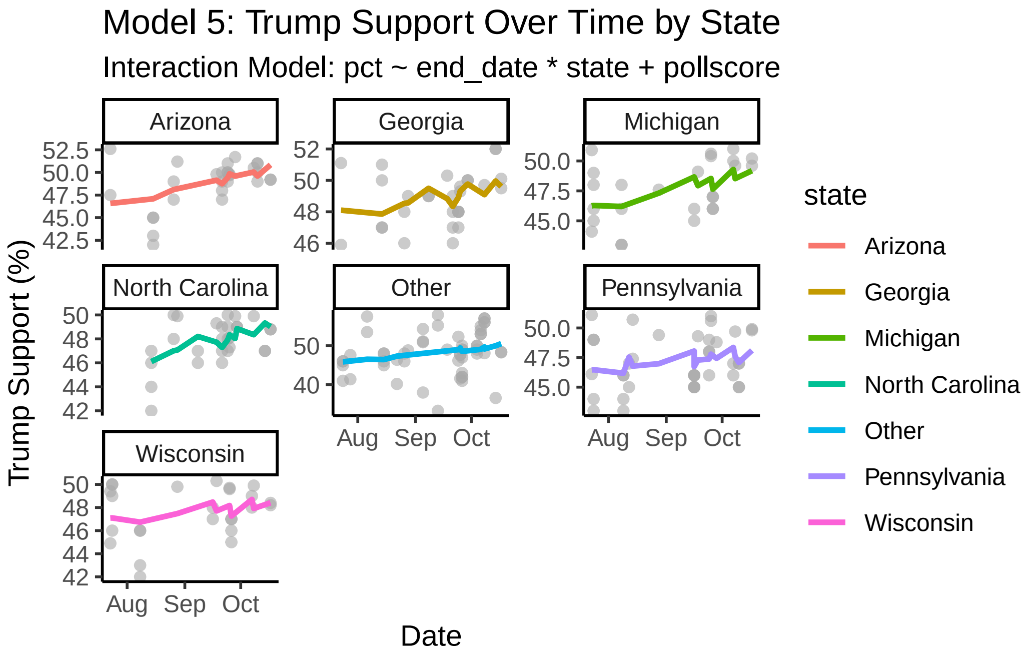

Figure 6: Trump Support Over Time by State, including Arizona, Georgia, Michigan, North Carolina, Pennsylvania, Wisconsin, and an aggregated "Other" category.

Figure Figure 7 illustrates the trend in Donald Trump's support percentage over time based on an interaction model that includes the predictors end_date, state, and pollscore. Each gray dot represents a poll observation, while the blue line indicates the smoothed trend in support across the polling period from August to October 2024.

The trend line suggests that Trump's support remained relatively stable, consistently hovering around the 50% mark. Despite this overall steadiness, individual poll results exhibit considerable variability, with support percentages ranging from below 40% to above 50% in some cases. This spread likely reflects differences in polling methodologies, sample demographics, or timing across individual polls.

The interaction model used (pct ~ end_date * state + pollscore) captures these nuances by allowing each state's support trend to vary over time while accounting for poll quality differences. This approach provides a more detailed view of how Trump's support might differ by region and according to the reliability of the polls.

In summary, while Trump's support shows minor fluctuations in specific polls, the overall trend indicates a stable base of support during this period. The dispersion in individual poll results underscores the importance of accounting for methodology and regional factors when interpreting polling data.

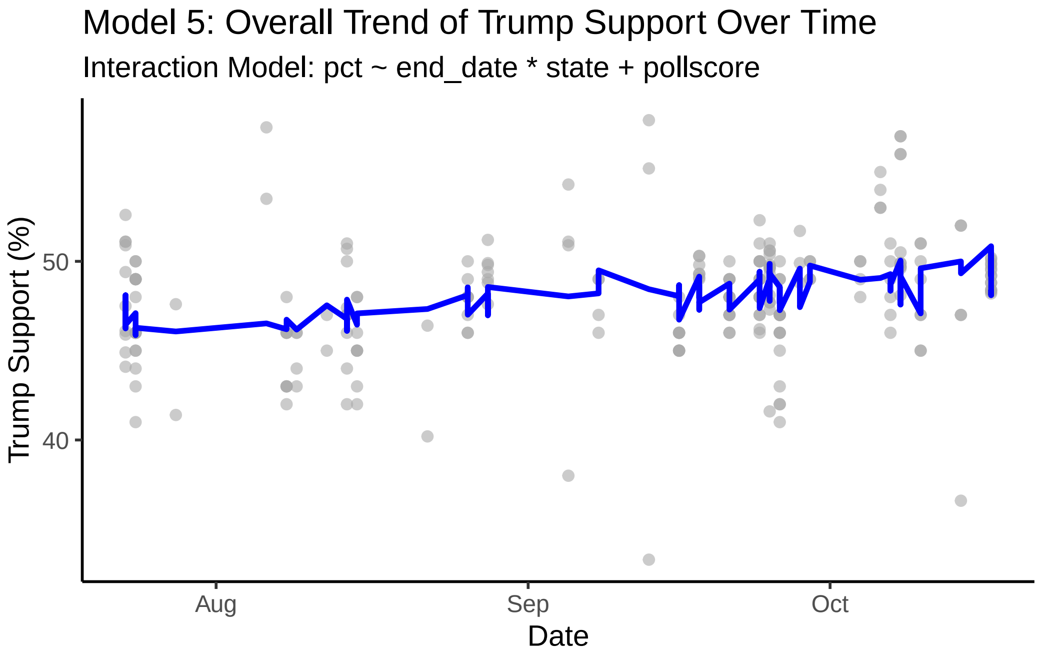

Figure 7: Overall trend of Trump support (%) over time based on an interaction model with the predictors end_date, state, and pollscore.

Figure Figure 8 shows the trend of Donald Trump's support over time in key battleground states, including Arizona, Georgia, Michigan, North Carolina, Pennsylvania, Wisconsin, and a combined "Other" category. The analysis utilizes an interaction model (pct ~ end_date * state + pollscore), with each state's trend line reflecting Trump's support percentage. The color gradient, ranging from blue to red, represents the poll score, with blue indicating lower quality polls and red indicating higher quality.

The trend lines in most states demonstrate a steady or slightly upward trend in Trump's support, particularly in Michigan and Georgia, where a gradual increase is observed. Meanwhile, North Carolina and Wisconsin show more variability, suggesting fluctuating levels of support. These patterns might reflect state-specific events or demographic differences that influence Trump's popularity.

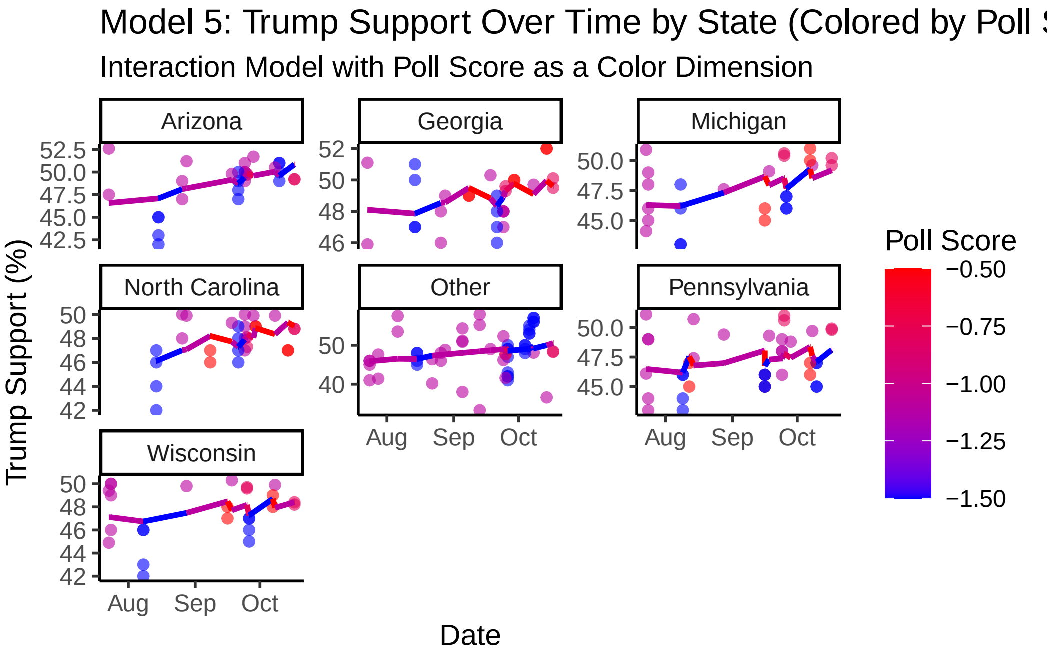

Figure 8: State-specific trends of Trump support (%) over time, with each line representing a different state and colors indicating poll quality. Warmer colors (red) represent higher poll scores, while cooler colors (blue) represent lower poll scores. This figure highlights regional variations and the influence of poll quality on observed trends in key battleground states.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Evaluation of YouGov's Polling Methodology

A.1 Overview

YouGov is a leading global polling organization known for its digital approach to data collection, utilizing an extensive online panel to conduct surveys across various domains, including politics. This particular survey, conducted from October 19-22, 2024, sampled 1,615 U.S. adults to assess voter preferences in the 2024 U.S. Presidential Election, focusing on support levels for Kamala Harris and Donald Trump. This appendix explores YouGov's online-only polling methodology, highlighting its core design, strengths, and limitations (YouGov, 2024a).

A.2 Population, Frame, and Sample of the Poll

- Target Population: YouGov's target population for this poll includes eligible U.S. voters, specifically adults likely to participate in the upcoming election. By focusing on likely voters rather than the broader adult population, YouGov increases the relevance and predictive power of its findings for the election (Groves et al., 2009).

- Sampling Frame: YouGov uses its proprietary online panel of over 24 million members globally as the sampling frame. To enhance the representativeness of its U.S. polling, YouGov applies demographic quotas based on key characteristics such as age, gender, and region, reflecting the general U.S. electorate (YouGov, 2024b).

- Sample: In this survey, YouGov selected 1,615 panel members from its online sampling frame. Although participation is voluntary, YouGov's use of demographic quotas and post-stratification weighting helps ensure that the sample reflects the U.S. adult voting population's diversity (Bethlehem, 2010).

A.3 Sampling Approach and Recruitment

YouGov recruits its panel members through various online channels, including social media advertisements and partnerships with third-party platforms, creating a diverse base of voluntary participants incentivized through redeemable points. Using stratified sampling, YouGov selects respondents based on demographic quotas that reflect the broader population, helping to mitigate some biases associated with online samples. However, the online-only sampling approach does exclude individuals without internet access, potentially leading to underrepresentation of certain groups, such as rural and elderly populations (Couper, 2000). While this approach is cost-effective and allows rapid data collection, the reliance on voluntary online participants introduces self-selection bias, as panel members may differ from the general population in terms of digital engagement and political awareness (Groves et al., 2009).

A.4 Advantages and Trade-Offs in Online-Only Sampling

Advantages:
The online-only approach used by YouGov offers significant advantages, particularly in terms of cost efficiency, data collection speed, and reach. Compared to traditional methods like phone or in-person interviews, online surveys are much less expensive, enabling YouGov to conduct regular polling cycles without prohibitive costs (Dillman, 2007). Additionally, online surveys allow for near-real-time data collection, which is invaluable in the fast-paced context of election cycles where public opinion can shift rapidly. With its extensive online panel, YouGov is also able to achieve substantial sample sizes, thereby enhancing the statistical power and reliability of its findings (YouGov, 2024b).

Trade-Offs:
The online-only approach used by YouGov introduces certain biases, particularly coverage and self-selection biases. Relying solely on internet-based surveys may exclude individuals without reliable internet access, which can create representation gaps, especially among lower-income or rural populations who may have limited connectivity (Bethlehem, 2010). Additionally, because participation in YouGov's panel is voluntary, there is a risk of self-selection bias; these participants may differ from the general population in meaningful ways, such as having greater engagement in political issues or higher digital literacy, which could influence their responses (Groves et al., 2009).

A.5 Non-Response Handling

YouGov applies post-stratification weighting to address non-response bias. Demographic weights based on age, gender, race, education, and political affiliation adjust the responses to align more closely with the U.S. population structure (Groves et al., 2009). By correcting for imbalances in response rates across different groups, this weighting approach aims to produce data that is more reflective of the overall electorate.

A.6 Questionnaire Design

Strengths: YouGov's questionnaire design emphasizes clarity, directness, and relevance, making it easier for respondents to provide accurate answers. The questions are designed to be straightforward and unbiased, ensuring that respondents can understand and respond without confusion, which enhances data quality (Dillman, 2007). Additionally, the survey is focused on core electoral topics, such as voter preferences and demographic trends, providing relevant and actionable insights that are especially valuable during election periods (YouGov, 2024a).

Weaknesses:
While YouGov's questionnaire is effective in capturing general voter preferences, it has certain limitations in depth and format. The questions do not explore the underlying motivations behind respondents' choices, which could offer deeper insights into voter behavior and the factors driving their preferences (Couper, 2000). Additionally, the purely online format restricts the potential for nuanced responses on complex issues, as respondents may provide less detailed answers than they would in a more interactive setting, such as in-depth interviews (Bethlehem, 2010).

A.7 Questionnaire Design

YouGov's methodology is well-suited for large-scale, cost-effective polling and is particularly advantageous for monitoring public opinion trends. Its use of demographic quotas and advanced weighting techniques enhances representativeness, but the online-only model presents challenges, including potential coverage and self-selection biases. These factors should be considered when interpreting YouGov's polling results.

B Idealized Methodology – Forecasting the 2024 U.S. Presidential Election

B.1 Overview

This methodology presents a detailed, budget-conscious approach to forecasting the 2024 U.S. presidential election. With a $100,000 budget, the plan combines a multi-modal sampling strategy, thorough data validation, and real-time data aggregation to capture a representative snapshot of voter sentiment. Drawing on best practices from established pollsters like TIPP Insights and Emerson College Polling, this methodology integrates online and phone-based sampling to ensure both cost-effectiveness and comprehensive demographic coverage (Emerson College Polling, 2024; TIPP Staff, 2023).

B.2 Sampling and Recruitment Strategy

The target population includes U.S. voters likely to participate in the 2024 election, with an objective to gather a sample of 9,000 respondents. A multi-modal approach maximizes reach across demographics:

Online Recruitment: To engage younger, urban, and digitally connected voters, pre-screened online panels from platforms such as Dynata are utilized. Social media ads and email invitations further expand online outreach, with each respondent receiving a $3 digital gift card for participation. This method effectively reaches digitally engaged voters and enhances sample diversity (Dillman, 2007).

Phone Sampling: To reach older and rural demographics less accessible online, live phone interviews are conducted using random digit dialing from verified voter lists, offering $5 gift cards as compensation. This approach ensures balanced inclusion of groups who may not participate in online surveys, improving overall sample representativeness (Bethlehem, 2010).

SMS Invitations: For mobile-first respondents, SMS invitations link to a mobile-optimized survey. This format is particularly effective in reaching younger voters who frequently use mobile devices, broadening the demographic base.
The sample is stratified by age, gender, race, education, and geographic region according to U.S. Census data, ensuring that the survey captures a balanced representation of the electorate (Elliott, 2020).

B.3 Data Collection and Validation

Data collection is centrally managed through Google Forms for online responses and phone interviews, enabling efficient quality control and secure handling. To ensure high data quality, several validation measures are integrated into the process. Duplicate checks monitor IP addresses and phone numbers to prevent multiple responses from the same individual, ensuring that each participant is counted only once. Consistency checks with embedded logic identify illogical or contradictory answers, filtering out low-quality data that could introduce bias into the final analysis (Groves et al., 2009). After data collection, post-stratification weighting is applied to align responses with U.S. Census demographics, adjusting for any over- or under-represented groups, which enhances both the reliability and representativeness of the results (Elliott, 2020). These steps ensure that the final dataset is accurate and aligned with population benchmarks, preserving the integrity of the forecast.

B.4 Data Aggregation and Analysis

A rolling aggregation method continuously updates and re-weights incoming data, capturing real-time shifts in voter sentiment. As Election Day approaches, the sample size will increase to capture any last-minute changes in public opinion, improving the model's responsiveness. Adjustments for undecided voters will be based on historical trends, enhancing the forecast's accuracy by accounting for late-deciding segments of the electorate (Bethlehem, 2010; Dillman, 2007).

B.5 Conclusion

By integrating a balanced sampling approach, rigorous validation techniques, and rolling data aggregation, this methodology provides a robust, cost-effective model for forecasting the 2024 U.S. presidential election. Its emphasis on demographic diversity, real-time adaptability, and data accuracy makes it well-suited for predicting voter behavior. This structured approach combines the strengths of established methodologies and adapts them for a comprehensive and insightful election forecast.

A link to the survey can be found at: https://forms.gle/15J7dTGn8iELasJ68

C Copy of the Survey

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

C.1 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

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