Modeling 2024 U.S. Presidential Polling Trends*

Democrats Show Strong Support Leading Up to the Election, with Biden's Steady Backing and Harris's Fluctuations Across Regions and Polls.

Mandy He Wendy Yuan

November 4, 2024

This paper looks at polling data from the 2024 U.S. Presidential election to see how candidate support changes and what factors affect polling results. We found that Biden's support remained steady, while Harris's numbers fluctuated, especially in certain states and under specific pollsters. These finding shows that the Democrats appear to be in a strong position heading into the election. Used alongside election predictions, these insights help clarify the broader scope of Democratic support leading up to the election.

Table of contents

*Code and data are available at: https://github.com/MandyHe7/US-Election.git.

1 Introduction

The 2024 U.S. Presidential election is a major event in American politics, with polling data playing a vital role in shaping public opinion and campaign strategies. Polls are used to measure public sentiment and predict election outcomes, but the reliability of these predictions depends on factors such as polling methodology and geographic focus. Although there is a large amount of polling data available, understanding how these factors affect the accuracy of predictions remains a challenge.

This paper does not attempt to forecast the election outcome. Instead, it examines trends in polling support over time and analyzes the factors that contribute to variations in candidate support. Using data from FiveThirtyEight's U.S. Presidential election polls, collected from pollsters such as YouGov and Siena/NYT, we model polling percentages for each candidate, considering variables like pollster and region. By exploring these variables, we aim to enhance our understanding of how they affect polling results.

We focus on analyzing trends in polling data rather than forecasting the election outcome to gain a deeper understanding of how and why public opinion changes over time. By examining trends, we can see how specific events, regions, and campaign efforts impact voter preferences, providing insights that pure forecasting might miss. Trend analysis also avoids the need for

assumptions about future behavior, giving us a clearer picture of the factors that shape current voter support. This approach helps campaigns and analysts recognize consistent patterns in support, guiding more targeted strategies without relying on uncertain predictions.

The results show clear trends in candidate support leading up to the election, with differences based on pollster methods, geography, and timing. These findings provide valuable insights into the dynamics of public opinion during the election period, rather than making predictions about the outcome.

The remainder of the paper is structured as follows: Section 2 details the data and measurement process. Section 3 describes the methodology, including linear and Bayesian models. Section 4 presents the results, highlighting trends in polling, and Section 5 considers the implications of these results for future research on polling and public opinion.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to analyze polling data sourced from FiveThirtyEight's U.S. Presidential election polls (FiveThirtyEight 2024), collected by organizations like YouGov and Siena/NYT through online surveys and phone interviews. The dataset includes details such as pollster, methodology, sample size, and location, allowing us to track how public opinion shifts over time and across regions. After cleaning the data for accuracy, we analyze trends in candidate support using the outcome variable pct (polling percentage) and predictor variables like pollster and state to model and forecast public opinion trends as the election approaches.

2.2 Measurement

Our data represents real-world public opinion on the 2024 U.S. Presidential election, gathered through polls conducted by various organizations. These polls ask people which candidate they plan to support, using methods like online surveys or phone interviews. This process starts with asking a group of people, typically selected to represent the larger population, about their voting preferences.

The information gathered from these people—such as who they support, where they are located, and when the poll was conducted—is then recorded in a structured format. For example, in a YouGov poll conducted in Arizona from October 11 to 16, 2024, a percentage of respondents supported Harris. These real-world responses become data points in our dataset, capturing the details of the poll, including the pollster, methodology, sample size, and dates.

Once this data is collected, it's cleaned to ensure accuracy by selecting numerical grade that is equal to 3, removing any inconsistencies or missing values. This allows us to analyze and track

how public opinion changes over time and in different regions, turning individual responses into a clear and usable dataset.

2.3 Outcome variables

The primary outcome variable in this dataset is pct, representing the percentage of respondents in a particular poll who support a specific candidate. This variable is the central focus of our analysis, as it reflects the level of public support for each candidate at the time the poll was conducted. Additionally, the variable num_DEM captures the number of Democratic respondents in each sample, providing context on party-based support patterns.

2.4 Predictor variables

The following predictor variables explain variations in polling outcomes:

- pollster: Identifies the polling organization (e.g., YouGov, Siena/NYT). Different pollsters may use varied methodologies, which can introduce variability in results.
- state: Indicates the region or national focus of the poll, capturing geographical differences in voter behavior.
- end_date: Capture the timeframe in which the poll was conducted, enabling the analysis of trends over time.

These variables allow us to model polling percentages and forecast trends, taking into account factors such as pollster and regional variations. A summary statistics for predictor variables can be found in Section C.

3 Methodology

3.1 Data Collection and Preprocessing

We used R (R Core Team 2023) to process polling data for the 2024 U.S. Presidential Election from (FiveThirtyEight 2024). The data includes results across demographic groups and time periods. Missing values were addressed, and dummy variables created for categorical predictors like region and pollster. Post-stratification weighting ensured the sample represented the U.S. population, reducing bias.

3.2 Model Selection

We tested various regression models to predict candidate vote percentages, using predictors like polling date, candidate, and pollster. Models ranged from simple linear regressions to Bayesian models, incorporating historical election data through Bayesian priors. We chose to use simple linear regression and Bayesian models instead of multiple regression due to the low R-squared values obtained from the multiple regression analysis, which indicated that they do not provide significant insights see Section D.

3.3 Simple Linear Regression

We are performing a regression analysis of PCT with respect to the end date. The following is the simple linear regression equation $pct = \alpha + \beta_1 \cdot \text{end\_date} + \epsilon$.

3.4 Bayesian Modeling

For the Bayesian model, MCMC sampling estimated parameters, providing credible intervals. The model updated dynamically with new polling data, improving forecasts over time.

3.5 Assumptions and Limitations

We assumed a linear relationship between predictors and outcomes. Although Bayesian methods helped capture non-linearities, some complexities may remain. Non-response bias could still affect results despite weighting.

3.6 Software and Tools

We use the statistical programming language R (R Core Team 2023) and the following packages: tidyverse (Wickham et al. 2016), dplyr (Wickham 2023a), readr (Wickham 2023c), ggplot2 (Wickham 2023b), janitor (Baumer et al. 2023), lubridate (Wickham et al. 2023), broom (Robinson et al. 2023), modelsummary (P. 2023), rstanarm (Goodrich et al. 2023), and splines (Team 2023), knitr (Xie 2023), kableExtra (Huang 2023).

4 Results

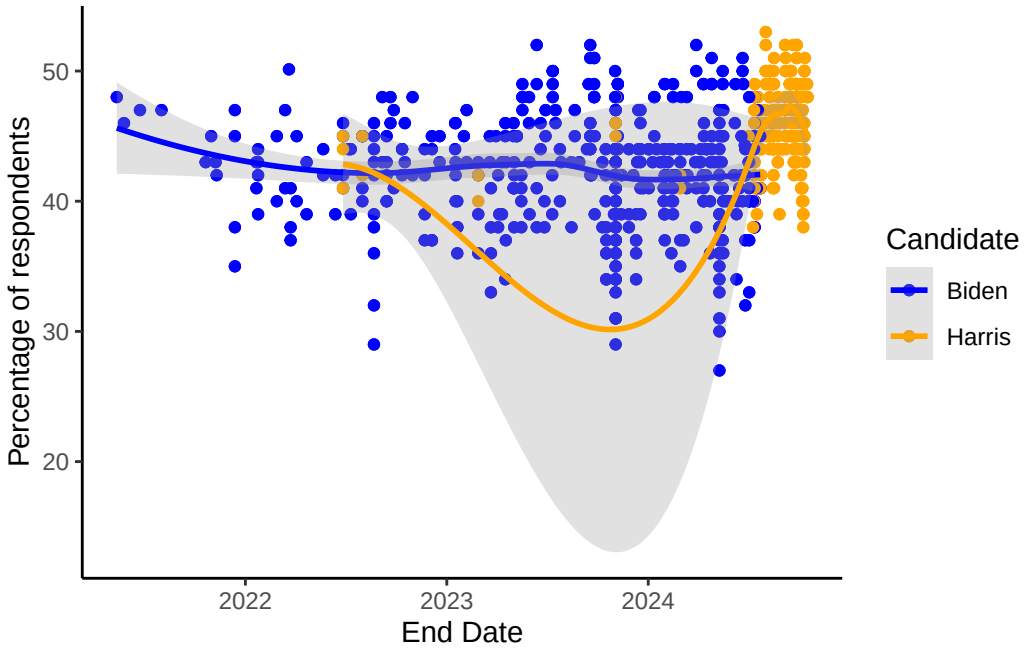

Figure 1: Simple linear regression between end date and precentage of respondent votes, showing a positive trend for Democrats with a noticeable dip in the middle.

Figure 1 illustrates a simple linear regression between end date and pct. It shows a positive trend between the variable but with a noticeable dip around the middle. This dip may indicate some underlying complexity that the simple linear model doesn't fully capture, but the model does a reasonable job fitting the general positive direction of the data. The regression line captures this linear relationship, with most data points clustering near the line, suggesting a moderate fit. The confidence interval shows some uncertainty at extreme values, where the data becomes more dispersed.

Figure 2 shows a linear regression applied to multiple subsets of the data. Each panel represents a different pollsters and the regression line within each panel shows a generally consistent positive trend across most subsets - although there are some dips in the trend. Some subsets, like YouGov, have more data points than others, leading to more robust lines with narrower confidence intervals. The individual panels help highlight how the relationship between end date and pct across different pollster, revealing subtle variations between groups.

Figure 3 apply a linear regression across various states. Each panel depicts a clear linear relationship within its subset, with a up and down wave trend in most. Some panels, particularly those with more data points, show stronger, clearer lines, while others may have flatter trends or greater variability due to smaller sample sizes. This provides a clear comparison

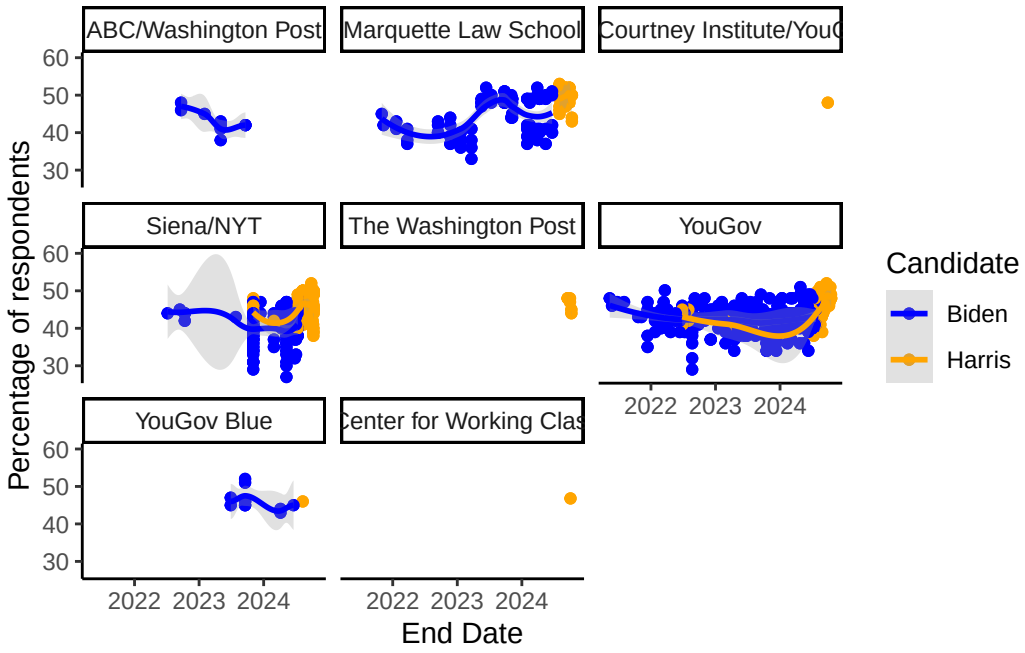

Figure 2: Linear regression applied to multiple subsets of the data. Each panel represents a different pollster, with regression lines indicating a generally consistent positive trend.

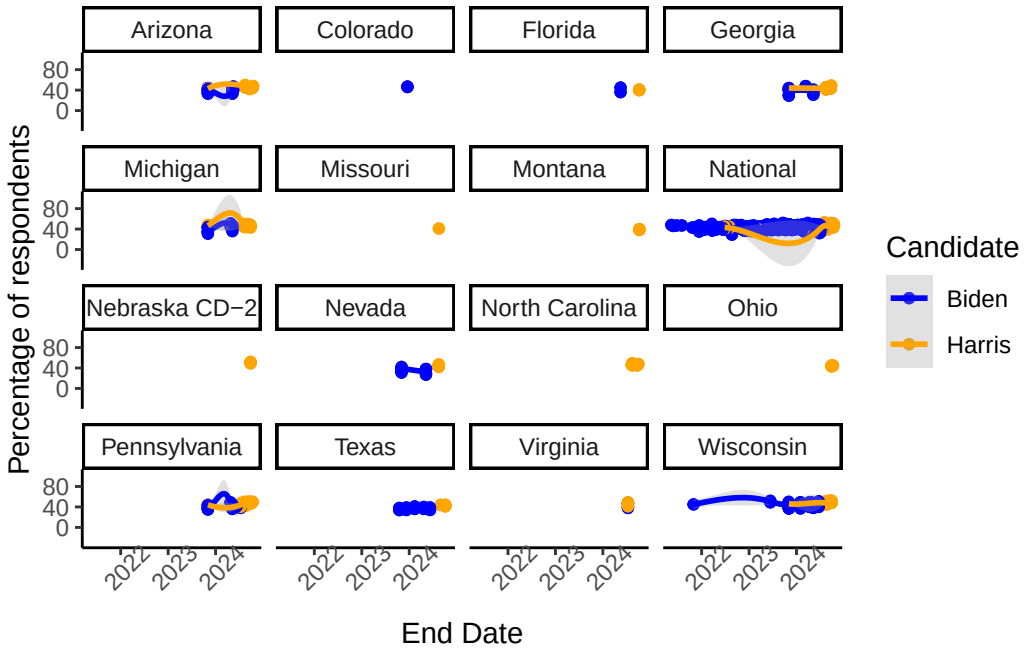

Figure 3: Linear regression across various states. Each panel reveals a linear relationship with clearer trends.

across groups, making it easy to see where the linear relationship holds strongly and where it varies.

The Bayesian model in Figure 4 shows a moderate upward trend where pct increases as end date rises. The fitted line follows a generally smooth, linear shape, though there are slight fluctuations around the line. The credible intervals are wider at the more recent year values of end date, indicating increased uncertainty. The overall shape is slightly curved, but still retains a mostly linear form, with the credible intervals giving the graph a distinctive funnel shape—wider at the edges where the data is sparser, and narrower in the middle where data points are concentrated.

Figure 5 in a Bayesian model that shows different colour points each corresponding to different states. The overall trend across most groups is an upward slope, though some groups have flatter or even slightly downward trends. The lines vary in steepness, and the credible intervals for each group give the plot a multi-layered shape, with each group showing distinct levels of uncertainty. The shape of the data suggests that many groups follow a similar positive trend. The graph as a whole has a spread-out, fan-like appearance, with a spreading out across the plot, reflecting the varied behaviors.

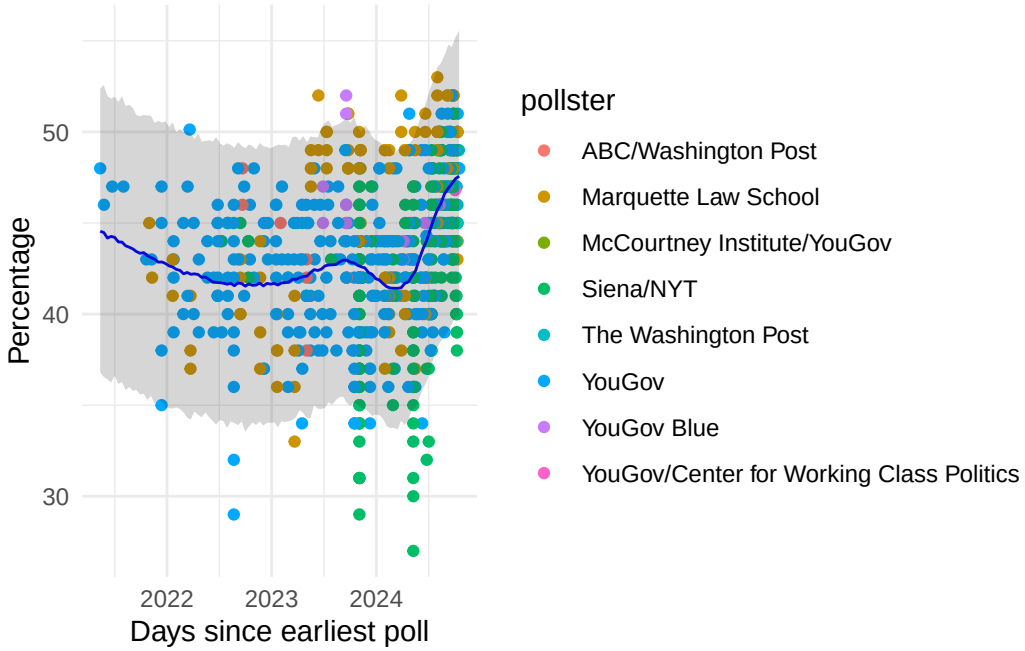

Figure 4: Bayesian model using spline fit to show a moderate upward trend, where percentage increases over time.

5 Discussion

5.1 Analyzing Polling Trends Over Time: Comparing Biden and Harris

Figure 1 shows the polling trends for Biden (blue) and Harris (orange) over time using a linear model with spline fit. Biden's polling starts above 45%, gradually declines, and then stabilizes toward 2024. His trendline is relatively flat, with narrow confidence intervals, indicating stable support and less variability in the data.

Harris shows a different pattern, with her polling starting around 40%, dipping significantly mid-2023, and then rising sharply into 2024. The wider confidence intervals around her trend, especially during the dip and rise, reflect greater uncertainty and fluctuation in her polling.

This model effectively captures the key trends: Biden's steady support versus Harris's more volatile polling. The model highlights the differences in public opinion dynamics for each candidate during the given period.

Furthermore, Harris's polling fluctuations reflect Keyes's (Keyes 2019) idea of "data violence," which suggests that standard polling methods may unintentionally reinforce narrow or inconsistent views across different demographic groups. Keyes's focus on using diverse data sources

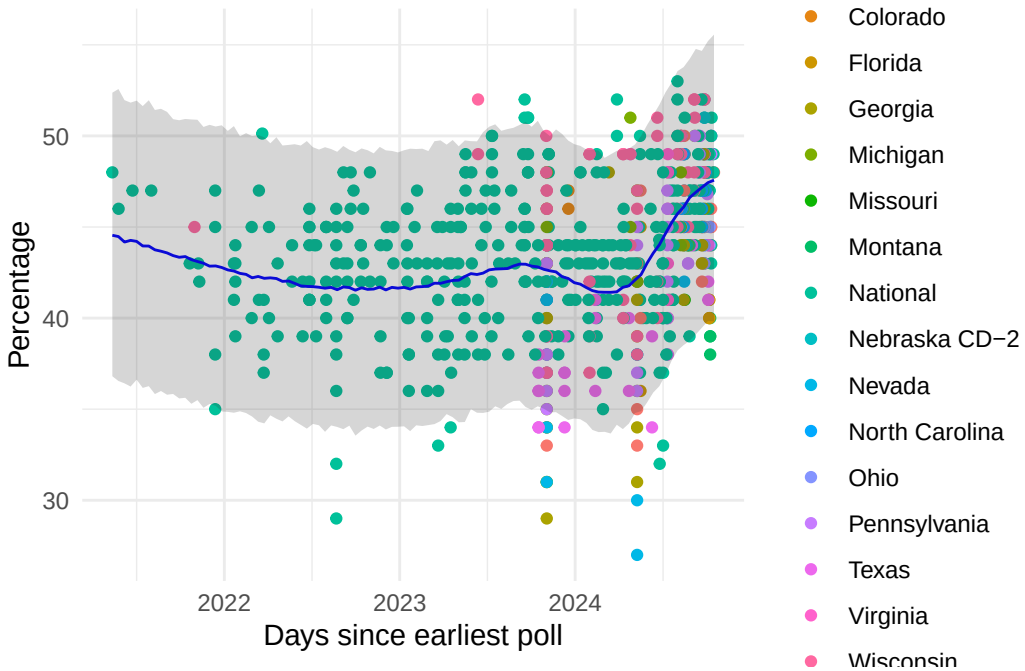

Figure 5: Bayesian model using spline fit with color-coded points for different states showing a moderate upward trend.

could help reduce these biases and offer a more accurate picture of candidate support among various population segments.

5.2 Analyzing Pollster-Specific Trends in Candidate Support

Figure 2 uses a linear model to illustrate the polling trends for Biden (blue) and Harris (orange) across different pollsters. Each panel represents the polling data from a specific pollster, showing how candidate support varied over time from 2022 to 2024.

For most pollsters, Biden shows a relatively stable trend with minor fluctuations, as seen in panels like ABC/Washington Post and YouGov Blue. However, in the Marquette Law School and YouGov panels, Biden's polling follows a slight upward trend toward the end, indicating a potential rise in support over time. Harris's trend is more variable, particularly in pollsters like Siena/NYT and YouGov, where her polling dips mid-period but recovers sharply as 2024 approaches.

The shaded regions represent confidence intervals, with wider intervals in some panels reflecting greater uncertainty in those pollster's results. For instance, Siena/NYT has notably larger confidence intervals, signaling more variability in the data compared to YouGov, where the intervals are narrower, indicating more consistent polling results.

Figure 2 highlights how different pollsters capture varying dynamics in candidate support. While the overall trend for Biden remains more stable, Harris's support appears to fluctuate more significantly, with clear differences in polling outcomes between certain pollsters. These variations emphasize the importance of considering pollster-specific methodologies when interpreting election forecasts.

5.3 Analyzing State-by-State Polling Trends in Candidate Support

Figure 3 shows the polling trends for Biden (blue) and Harris (orange) across various U.S. states using a linear model. Each panel represents a separate state or district, allowing for comparison of candidate support trends from 2022 to 2024. The National panel includes polls that do not have state-specific data, providing a broader, generalized view of national polling trends.

In most states, Biden's polling is relatively stable, as seen in Arizona, Michigan, and Wisconsin. He maintains consistent support, with narrow confidence intervals indicating lower uncertainty. In Ohio and National, Biden experiences a dip around 2023, but recovers toward 2024, mirroring broader national trends.

Harris's support is more variable across states. In states like Georgia and Nebraska CD-2, her polling is sparse, and in some cases, no clear trend is discernible. However, in Michigan and Wisconsin, Harris's polling tracks closely with Biden's. In the National panel, Harris shows a significant dip in 2023, followed by a sharp recovery leading into 2024, which is consistent with the trends seen in other key states like Ohio.

The National panel helps contextualize overall candidate trends by showing the aggregate polling data, without specific state breakdowns. This allows us to see general patterns in polling, though it doesn't capture state-specific variations, which are crucial for understanding electoral dynamics.

This analysis highlights Biden's more stable polling trends across most states, while Harris's polling exhibits more fluctuation, with the National data serving as a useful broad comparison for state-level trends.

5.4 Analyzing Bayesian Polling Trends of Different Pollsters

Figure 4 presents polling percentages over time, color-coded by different pollsters, using a Bayesian model with a spline fit. The key feature of the Bayesian model is the incorporation of uncertainty directly into the analysis, represented by the shaded credible interval. This interval helps capture the range of likely polling percentages, accounting for both the variability in the data and differences across pollsters.

The spline fit shows a general decline in polling percentages up until mid-2023, followed by a sharp rise as the election approaches in 2024. The credible intervals are wider earlier on,

indicating greater uncertainty due to less consistent polling data, and narrow closer to the election, reflecting increased confidence as more data is collected.

The color coding by pollster highlights how different organizations contribute to this trend. Pollsters such as YouGov and The Washington Post provide a larger volume of data points over the timeline, which helps tighten the credible intervals. Pollsters with fewer data points, like YouGov Blue or Siena/NYT, contribute less, leading to more uncertainty in their sections of the plot.

The Bayesian model is particularly important here as it explicitly models the uncertainty, allowing us to see where our predictions are more or less certain. This is crucial when comparing data from multiple pollsters, as some provide more consistent data than others. By incorporating this uncertainty, the model gives a clearer, more reliable picture of overall polling trends.

5.5 Analyzing Bayesian Polling Trends Differentiated By States

Figure 5 presents polling percentages over time, separated by states, using a Bayesian model with a spline fit. Each state is represented by a different color, providing a clear comparison of state-specific polling trends. The blue spline line shows the overall trend across all states, and the shaded area represents the credible interval, capturing the uncertainty in the predictions.

The Bayesian spline model reveals a decline in polling percentages leading up to 2023, followed by a notable increase approaching 2024. The credible interval is wider earlier on, reflecting greater uncertainty in polling results, and narrows as we move closer to the election, indicating more confidence in the predictions as more data becomes available.

The data points from various states, such as Michigan (green) and Ohio (light blue), cluster around the overall trend, contributing significantly to the model's fit. States with fewer data points, like Virginia (pink) and Nebraska CD-2 (light blue), contribute less and have higher variability, adding to the overall uncertainty in those periods.

This figure highlights the importance of incorporating state-level polling into the analysis, as each state has a different polling trajectory. The Bayesian model's ability to incorporate uncertainty provides a more reliable view of the overall trend, while still allowing for variability between states.

5.6 Weaknesses and next steps

A key weakness in the analysis is the variation between pollsters. Despite adjustments and the use of Bayesian modeling, differences in methodologies, sampling techniques, and respondent recruitment remain a challenge. Some pollsters, particularly those with fewer data points, contribute to greater uncertainty, which can make predictions less reliable in certain cases.

Additionally, relying on self-reported polling data carries the risk of biases, such as non-response bias or unrepresentative samples.

The use of spline models to capture non-linear trends is also a limitation. While it smooths fluctuations, it may overlook significant events or shifts in public opinion during the election period. Although the Bayesian model helps account for uncertainty, it relies on historical data for priors, which may not fully reflect the unique aspects of the 2024 election.

For future improvements, incorporating more detailed data, such as voter demographics and turnout probabilities, could enhance the accuracy of the forecasts. Integrating real-time polling updates or analyzing social media sentiment could make the models more responsive to changing conditions. Expanding the analysis to include more pollsters and considering the impact of electoral college dynamics would provide a more complete understanding of election outcomes beyond the popular vote.

To improve our paper, we can draw on the insights of Neyman (Neyman 1934) and Keyes (Keyes 2019). Neyman's work highlights the value of stratified sampling, which could guide us in making samples more representative across demographic groups, ultimately strengthening our forecasts. We could incorporate voter demographics and turnout probabilities data to enhance our model's accuracy. Keyes emphasizes the ethical importance of using diverse data sources. Applying this to our polling methods would help ensure fair representation across various demographics and viewpoints. Finally, expanding the analysis to include more pollsters and accounting for the electoral college's role would provide a more comprehensive view of election outcomes, not limited to the popular vote alone.

6 Conclusion

This paper analyzed polling data for the 2024 U.S. Presidential election, focusing on trends in candidate support and the factors that influence polling outcomes. Using data from FiveThirtyEight, which compiles polls from organizations like YouGov and Siena/NYT, we modeled polling percentages across different timeframes, pollsters, and states. Both linear regression and Bayesian models were applied to capture variations in the data and estimate the uncertainty in polling predictions.

The analysis showed that Biden's support remained fairly stable throughout the election period, while Harris's polling was more variable, with notable fluctuations over time and across pollsters. State-specific trends demonstrated the importance of regional differences in voter preferences, with certain states showing stronger or weaker support for each candidate. Bayesian models provided a way to quantify uncertainty, particularly for states or pollsters with fewer data points.

While the analysis provides useful trends, it is limited by pollster variability and the potential for bias in self-reported polling data. Future work should include more detailed voter demographics and real-time polling updates to enhance the accuracy of forecasts.

Based on the current trends in the data, the Democratic Party shows stable support, especially for Biden, heading into the election. Although there are fluctuations in Harris's polling, the overall trend suggests that the Democrats are in a relatively strong position. However, continued monitoring of polling trends, particularly at the state level, will be important for a more precise forecast of the election outcome.

A Appendix A: YouGov Pollster Methodology and Evaluation

Overview of YouGov

YouGov is a widely recognized international polling firm known for its political forecasting, especially through its "poll-of-polls" methodology, which aggregates results from numerous individual surveys(YouGov 2023a). YouGov primarily utilizes online panel polling, offering timely and cost-effective results, especially for U.S. presidential elections. According to Cirone and Spirling's (Cirone and Spirling 2021) concept of "methodological awareness" is important here, as understanding how each pollster's methods affect their results can help create a clearer view of national support trends. Below is a detailed exploration of YouGov's polling methodology (YouGov 2023b).

Population, Frame, and Sample

- Population: YouGov's target population consists of U.S. likely voters, defined by their eligibility and likelihood to vote, assessed through past voting behavior, voter registration, and expressed intent.
- Frame: The sampling frame is YouGov's online panel, which includes millions of U.S. residents recruited through various methods. The panel is stratified and weighted to represent key demographics such as age, gender, race, education, and region.
- Sample: YouGov uses non-probability sampling from its panel, adjusting for biases using sophisticated weighting to reflect the U.S. electorate's composition.

Sample Recruitment

- YouGov recruits panelists through:
  - Targeted online ads: Ads across websites and social media attract diverse participants.
  - Affiliate marketing: Collaborations with other websites help reach a broad audience.
  - Incentives: Respondents earn points for surveys, increasing participation and retention.

However, because respondents voluntarily join the panel, there is potential for self-selection bias, as certain demographics may be more likely to participate.

Sampling Approach and Trade-offs

YouGov applies stratified sampling, setting quotas for demographics such as age, gender, race, education, and region. After data collection, post-stratification weighting adjusts for imbalances between the sample and the population.

- Advantages:
  - Cost-effectiveness: Online panels are cheaper than traditional phone surveys.

  - Speed: Surveys are completed quickly online.
  - Targeting: Specific groups, such as likely voters, can be efficiently targeted.

- Disadvantages:

  - Non-probability sample: Not all individuals have an equal chance of being selected, risking bias.
  - Internet access bias: Those without internet access may be underrepresented, skewing results, - particularly for older or lower-income demographics.

Handling Non-response

To combat non-response, YouGov:

- Incentivizes participation with rewards to increase survey completion.
- Sends follow-ups for longer surveys, reminding respondents to complete them.
- Applies weighting to adjust for differential non-response (e.g., giving more weight to younger voters if they are underrepresented).
- Despite these efforts, some non-response bias may persist, as those opting out may differ systematically from respondents (e.g., lower political engagement).

Questionnaire Design: Strengths and Weaknesses

- Strengths:

  - Clarity: Questions are designed for consistency in interpretation.
  - Pre-testing: Surveys are often tested with smaller groups to refine clarity and accuracy.
  - Flexibility: Questions are adapted to political events, maintaining relevance.

- Weaknesses:

  - Limited depth: Online surveys favor simpler, multiple-choice questions, limiting nuanced responses.
  - Response fatigue: Frequent survey participation may cause rushed or less thoughtful responses.
  - Exclusion of offline participants: Individuals without internet access are excluded, impacting representativeness, particularly in underserved demographics.

Conclusion

YouGov's online panel approach offers cost efficiency, speed, and adaptability, making it suitable for political forecasting. While its methodology effectively mitigates some biases through stratification and weighting, limitations like non-response bias, self-selection, and internet access gaps must be acknowledged. Despite these challenges, YouGov remains a leading firm in modern political forecasting.

B Appendix B: Idealized Methodology and Survey

Overview

With a $100K budget, the objective of this survey is to forecast the outcome of the upcoming U.S. presidential election using a robust sampling strategy, respondent recruitment process, and careful poll aggregation techniques. This methodology is designed to minimize bias and ensure a representative sample that accurately reflects the voting population.

Sampling Approach

The goal of the survey is to represent the U.S. electorate. To achieve this, the sampling strategy should include a multi-stage stratified random sampling approach. The following demographic groups should be stratified to ensure diversity:

- Age: Ensure a balanced representation across different age groups (e.g., 18-29, 30-44, 45-64, 65+).
- Gender: Include a balanced proportion of male and female respondents, and allow space for non-binary respondents.
- Race/Ethnicity: Ensure representation of major racial and ethnic groups, including White, Black, Hispanic/Latino, Asian, and other minorities.
- Education: Include respondents with various educational backgrounds (no high school, high school graduate, college graduate, post-graduate education).
- Region: Sample respondents from different geographic regions (Northeast, Midwest, South, West) to capture regional variations in voting behavior.
- Sample Size: Given the budget constraints and the need for statistical reliability, the survey would target a sample size of 5,000 respondents, which allows for meaningful subgroup analysis while keeping costs manageable. This should provide a margin of error around $\pm 1.5\%$ at a 95% confidence level.

Recruitment of Respondents

To recruit a diverse and representative sample, the following methods will be used:

- Online recruitment via targeted ads: Similar to YouGov, targeted ads on social media platforms (e.g., Facebook, Instagram, YouTube) and websites can be used to attract a broad range of respondents. Ads will be tailored to reach different demographic groups based on age, region, and other factors.
- Email outreach to existing voter databases: Leveraging voter registration lists and databases (where legally permissible) will allow for targeted invitations to participate in the survey.
- Incentives: Offering incentives, such as entry into a lottery or digital rewards, will increase the participation rate and reduce non-response bias.

Given the budget, $80,000 will be allocated for recruitment efforts and respondent compensation, including targeted online ads and incentives for participants.

Data Validation

To ensure the data's accuracy and reduce the risk of fraudulent responses or duplicate submissions, the following measures will be implemented:

- Captcha verification: This will prevent bots from filling out the survey.
- Email and phone verification: Respondents will be required to verify their email or phone number to ensure that each respondent is unique.
- Cross-verification with voter rolls: For respondents who voluntarily provide voter registration details, cross-checking with public voter rolls can validate their voting eligibility.
- Time tracking and completion rate monitoring: To detect inattentive or rushed responses, the time spent on each question will be tracked, and surveys completed unusually quickly will be flagged for further review.

Poll Aggregation

To provide a more robust forecast, this survey will be integrated into a broader poll-of-polls approach, aggregating data from multiple sources, including:

- Public polling data: In addition to this survey, data from other reputable pollsters (like YouGov, Ipsos, etc.) will be aggregated.
- Weighting and adjusting: Each poll's results will be weighted based on its sample size, methodological rigor, and recency. Larger, more recent, and methodologically sound polls will have a greater influence on the final aggregated forecast.
- Adjustments for known biases: If a particular demographic is underrepresented (e.g., younger voters, rural voters), post-stratification weighting will be applied to ensure the aggregate data reflects the overall voting population.

Budget Breakdown

- Recruitment (online ads and outreach): $60,000
- Incentives for respondents: $20,000
- Survey development and data validation: $10,000
- Poll aggregation and analysis tools: $10,000

Survey Implementation

To demonstrate the methodology, a Google Forms survey will be developed and deployed. The survey will be structured as follows:

- Introduction: A brief explanation of the survey's purpose, ensuring that respondents understand its relevance and importance.
- Demographic questions: These questions will collect essential data on the respondents' age, gender, race/ethnicity, education, and region.

- Political questions: The core section will ask about voting intent, candidate preference, party affiliation, and important political issues. These questions will be designed to minimize bias and allow for nuanced responses (e.g., Likert scales, multiple-choice options).
- Verification and consent: Respondents will be asked to verify their information and give consent for data use.

A link to the Google Forms survey will be included in the appendix, alongside a downloadable copy of the survey questions for transparency.

Survey Design Considerations

- Question clarity and neutrality: The questions will be carefully worded to avoid leading responses or introducing bias.
- Ordering of questions: Demographic questions will be asked first to set a neutral tone before moving into political questions.
- Pilot testing: Before full deployment, the survey will be tested on a small sample to ensure clarity and to refine any confusing questions.

Form link: https://forms.gle/epk9ptyVxZpVRVLq9

C Appendix C: Summary Statistics for Predictor Variables

pollster	n
YouGov	302
Siena/NYT	220
Marquette Law School	96
YouGov Blue	12
ABC/Washington Post	9
The Washington Post	6
McCourtney Institute/YouGov	1
YouGov/Center for Working Class Politics	1

Figure 6: Summary Statistics for Predictor Variables

state	n
National	383
Wisconsin	59
Pennsylvania	41
Arizona	29
Michigan	27
Georgia	24
Texas	24
Nevada	18
North Carolina	11
Virginia	8
Florida	6
Ohio	6
Montana	4
Nebraska CD-2	4
Colorado	2
Missouri	1

Figure 7: Summary Statistics for Predictor Variables

min_date	max_date	range_days
2021-05-13	2024-10-16	1252

Figure 8: Summary Statistics for Predictor Variables

D Appendix D: Multi-Linear Regression Result

This is a multiple linear regression that runs pct on end_date, pollster, and state. The regression equation is $pct = \alpha + \beta_1 \cdot \text{end\_date} + \beta_2 \cdot \text{pollster} + \beta_3 \cdot \text{state} + \epsilon$.

Statistic	Value
R-squared	0.2480289
Adjusted R-squared	0.2202675
F-statistic	8.9343171

Figure 9: Summary of Multi-Linear Regression: Key Statistics

References

Baumer, Ben et al. 2023. "Janitor: Simple Tools for Examining and Cleaning Dirty Data." https://cran.r-project.org/web/packages/janitor/index.html.

Cirone, Alexandra, and Arthur Spirling. 2021. "Turning History into Data: Data Collection, Measurement, and Inference in HPE." *Journal of Historical Political Economy* 1 (1): 127–54. https://doi.org/10.1561/115.00000005.

FiveThirtyEight. 2024. "2024 National Presidential Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Goodrich, Jonathan A. et al. 2023. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://cran.r-project.org/web/packages/rstanarm/index.html.

Huang, Yihui. 2023. "kableExtra: Construct Complex Table with 'Kable' and 'Markdown'." *The R Journal* 10 (1): 45–55. https://cran.r-project.org/package=kableExtra.

Keyes, Os. 2019. "Counting the Countless." *Real Life Mag.* https://reallifemag.com/counting-the-countless/.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. https://doi.org/10.2307/2342192.

P., Mikhael A. P. L. 2023. "Modelsummary: Create Summary Tables for Model Outputs." https://cran.r-project.org/web/packages/modelsummary/index.html.

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 et al. 2023. "Broom: Convert Statistical Analysis Objects into Tidy Tibbles." https://cran.r-project.org/web/packages/broom/index.html.

Team, R Core. 2023. "Splines: Regression Spline Functions and Basis Functions." https://cran.r-project.org/web/packages/splines/index.html.

Wickham, Hadley et al. 2016. "The Tidyverse." https://www.tidyverse.org.

Wickham, Hadley. 2023a. "Dplyr: A Grammar of Data Manipulation." https://cran.r-project.org/web/packages/dplyr/index.html.

———. 2023b. "Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics." https://cran.r-project.org/web/packages/ggplot2/index.html.

Wickham, Hadley et al. 2023. "Lubridate: Make Dealing with Dates a Little Easier." https://cran.r-project.org/web/packages/lubridate/index.html.

Wickham, Hadley. 2023c. "Readr: Read Rectangular Text Data." https://cran.r-project.org/web/packages/readr/index.html.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://cran.r-project.org/package=knitr.

YouGov. 2023a. "About YouGov." https://today.yougov.com/about.

———. 2023b. "Panel Methodology." https://today.yougov.com/about/panel-methodology.