Original Research Article

BAYESIAN MODEL FOR PREDICTING VOTING BEHAVIOR IN PRESIDENTIAL ELECTIONS IN KENYA

Abstract

This study applies the Bayesian Dirichlet-Multinomial model to predict Kenya's 2022 presidential election. The model incorporates voter age, gender, poverty rate, and political ideation as major determinants of voting behavior. The model estimated Raila Odinga's vote share at 48.17% (actual: 48.85%) and William Ruto's at 46.96% (actual: 50.49%). It slightly overestimated minor candidates but proved more accurate than polls. A Bayesian p-value of 0.487 confirmed model reliability. Bayesian inference demonstrated superior adaptability by quantifying uncertainty and updating probabilities. Key voter behavior determinants included youthful population and political ideation. The study concludes that Bayesian modeling enhances election forecasting and recommends integrating it with machine learning, social media sentiment analysis, and economic indicators for improved predictions.

Keywords: Presidential Elections, Election Forecasting, Bayesian Prediction Models

1 INTRODUCTION

Voting patterns and behavior are complex phenomena influenced by many factors, including voters 'demographic characteristics, attitudes and opinions on political issues, or their record of previous votes (Gherghina and Nemčok (2021)). Political campaigns need to identify the factors that predict voting behaviors. This will let them create messages to the voters more effectively and also be able to make informed decisions about resource allocation accordingly (Wang et al. (2015)).

In democratic societies, understanding and predicting voting behavior is paramount for political campaigns, policy makers, and researchers. Through the ballot, citizens express their preferences, and those elections determine paths that shape the future of their nation (Grover et al. (2019)). Kenya, one of the most multiethnic country in East Africa and vibrant politically, has had diverse voting patterns and ever-changing voter sentiments. Therefore, developing accurate models for predicting voting behavior in Kenya holds significant value for electoral strategies and the formulation of effective political policies.

In recent years, opinion polls in Kenya have been used to predict election outcomes, particularly presidential elections. Most of these polls correctly predicted the victory of the presidential candidate they supported. Some have also been used to gauge public support for candidates and parties and predict the election's outcome. Nevertheless, doubts have been raised about the reliability of these polls.

Kenyan politics, multi-ethnicity and regional differences are the basis with which Kenyan politics operate (Brass (2023)). This kind of diversity makes it difficult to understand why the electorate votes for a particular candidate and predict election outcome. There is, therefore, an increasing need for advanced models that can encompass in a complex way the interaction between the many factors affecting the voting behavior of its citizens in Kenya.

Despite the increasing use of opinion polls in Kenya to predict the outcome of presidential elections, the accuracy of these polls has been mixed. Some opinion polls have been successful in predicting the winner of the presidential race, while others have not. This raises the question of whether more sophisticated statistical models, such as Bayesian models, could be used to improve the accuracy of election predictions in Kenya.

The identified research gap revolves around the need for more comprehensive and dynamic Bayesian models that account for contextual and temporal dynamics, explore cross-context generalizability, and aim to predict individual-level voting behavior. By addressing these gaps, researchers can advance the field of Bayesian modeling for predicting voting behavior and contribute to more accurate and insightful predictions in the future.

The purpose of this study was to fit a Bayesian model for predicting presidential elections voting behavior in Kenya and to compare its performance to that of 2022 elections opinion polls. The study aimed to evaluate the Bayesian models in the context of predicting individual level voting behavior in presidential elections in Kenya.

2 Methodology

2.1 Model methodology

The Dirichlet multinomial regression model is a statistical modeling technique used for analyzing data with a categorical response variable and multiple independent variables. It is an extension of the multinomial regression model that takes into account the correlation among the response categories.

The Dirichlet distribution is a multivariate probability distribution defined on the simplex, which is a geometric space that represents all possible combinations of proportions or probabilities that sum to 1. It is parameterized by a vector of positive values, often denoted as , which controls the shape of the distribution.

Let Y be a categorical response variable with K categories ($K > 2$), and let $X = (X_i)$ where i=1,2 independent variables; ethnicity and political ideation (measured by political party affiliation of the voter). The response variable Y is represented as a K-dimensional vector, where each element Y_k represents the count or frequency of observations in category k.

The Dirichlet multinomial regression model assumes that the parameters of the multinomial distribution, denoted as $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$, follow a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$. The parameters $\boldsymbol{\alpha}$ determine the shape of the Dirichlet distribution and can be interpreted as prior counts for each category.

The relationship between the response variable Y and the independent variables X is modeled through a log-linear relationship. The logit of the category probabilities π_k is expressed as a linear combination of the predictors:

$$\text{logit}(\pi_k) = \beta_{0k} + \beta_{1k}X_1 + \beta_{2k}X_2 \tag{1}$$

where $\beta_{0k}, \beta_{1k}, \beta_{2k}$ are the regression coefficients of predictor variables; ethnicity and political ideation.

2.1.1 Posterior model

let Θ_{ij} be the true proportion of voters in a county that intend to vote the $\hat{\text{i}}$th candidate in a jth county. where i=1,2,3...k Candidates where k is the proportion of undecided voters and j=1,2...47 counties. The proportions of Θ_{ij} are assumed to be continuous in the interval [0,1] and sum to 1

The joint prior distribution for Θ_{ij} is assumed to be a conjugate prior distribution, thus, the resulting posterior belongs to the same distributional family as the prior distribution. To satisfy this requirement, we will assume Θ_{ij} follows a Dirichlet multinomial regression distribution $\theta_{ij}\tilde{}Dir\alpha$

$$\Theta_{ij} = \begin{pmatrix} \alpha_{11} & \alpha_{12}...\alpha_{147} \\ \cdot & \\ \cdot & \\ \alpha_{k1} & \alpha_{k2}...\alpha_{k47} \end{pmatrix}$$

which is a multivariate generalization of the beta distribution and is often used as a prior for the probability of a success in Bernoulli trials. Therefore, the joint probability density function can be written as

$$P(\Theta) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{n} \theta_i^{\alpha_i - 1} I(\theta \in S) \tag{2}$$

where $\Theta_{ij} > 0, i = 1, 2, 3...n, and \sum \theta_{ij} = 1$

The probability that a candidate wins a given county can be computed using the marginal probability densities. The joint probability density for the Dirichlet-multinomial model can be formulated as follows: The probability density function (pdf) of the Dirichlet distribution with parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ is given by:

$$f(p_1, p_2, p_3, p_4 \,|\, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = \frac{1}{B(\alpha_1, \alpha_2, \alpha_3, \alpha_4)} \cdot p_1^{\alpha_1 - 1} \cdot p_2^{\alpha_2 - 1} \cdot p_3^{\alpha_3 - 1} \cdot p_4^{\alpha_4 - 1} \tag{3}$$

where $B(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ is the multivariate beta function.

Assuming a logistic regression model for the relationship between the predictor variables (x_1, x_2) and the outcome probabilities (p_1, p_2, p_3, p_4), we have:

$$\log\left(\frac{p_1}{p_4}\right) = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 \tag{4}$$

where $\beta_1, \beta_2,$ and β_3 are the regression coefficients.

2.1.2 Joint Probability Density

The joint probability density of the Dirichlet-multinomial model can be obtained by multiplying the pdf of the Dirichlet distribution with the likelihood of the logistic regression model. Assuming independent observations, the joint probability density is given by:

$$f(p_1, p_2, p_3, p_4, x_1, x_2 \,|\, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3)$$
$$= \mathrm{f}(\mathrm{p}_1, p_2, p_3, p_4 \,|\, \alpha_1, \alpha_2, \alpha_3, \alpha_4) \cdot f(x_1, x_2 \,|\, p_1, p_2, p_3, p_4, \beta_1, \beta_2, \beta_3) \tag{5}$$

where $f(x_1, x_2 \,|\, p_1, p_2, p_3, p_4, \beta_1, \beta_2, \beta_3)$ is the likelihood function of the logistic regression model.

2.1.3 Data analysis

The Bayesian Dirichlet-Multinomial model was formulated and implemented using PyMC in Python, which provides functions for fitting probabilistic models through Markov Chain Monte Carlo (MCMC) sampling. The model estimated regression coefficients and hyperparameters of the Dirichlet distribution, capturing the relationship between the predictor variables and the vote distribution.

To assess model performance, we used posterior predictive checks (PPC) and Bayesian p-values to evaluate the goodness-of-fit. Additionally, convergence diagnostics such as R-hat values and Effective Sample Size (ESS) were examined to ensure reliable parameter estimation.

3 RESULTS AND FINDINGS

3.0.1 The Bayesian model

The model uses a set of independent variables, including socioeconomic indicators (poverty rate, voter turnout, gender ratio, youth percentage) and political affiliation (categorized as Azimio, Kenya Kwanza (KK), and Unflipped), to predict the vote shares. Based on the already determined hyper-parameters using the elbow method; $\alpha_1 = 2.0$, $\alpha_2 = 5.0$, $\alpha_3 = 1.0$, and $\alpha_4 = 3.0$ for Raila, Ruto, other candidates and undecided votes respectively.

The Bayesian model was run using the PyMC package in Python, implementing a Dirichlet-multinomial regression to estimate voting behavior in Kenya's 2022 presidential election. The model incorporated key predictors such as poverty rate, voter turnout, gender composition, youth percentage, and political ideation to determine their influence on candidate vote shares.

Markov Chain Monte Carlo (MCMC) sampling was used to estimate posterior distributions, with multiple chains run to ensure convergence and robustness. The model parameter tests, including posterior distributions and trace plots, are as shown in Figure 1 below. These diagnostics help assess the efficiency and stability of the sampling process, confirming whether the chains have reached a stationary distribution.

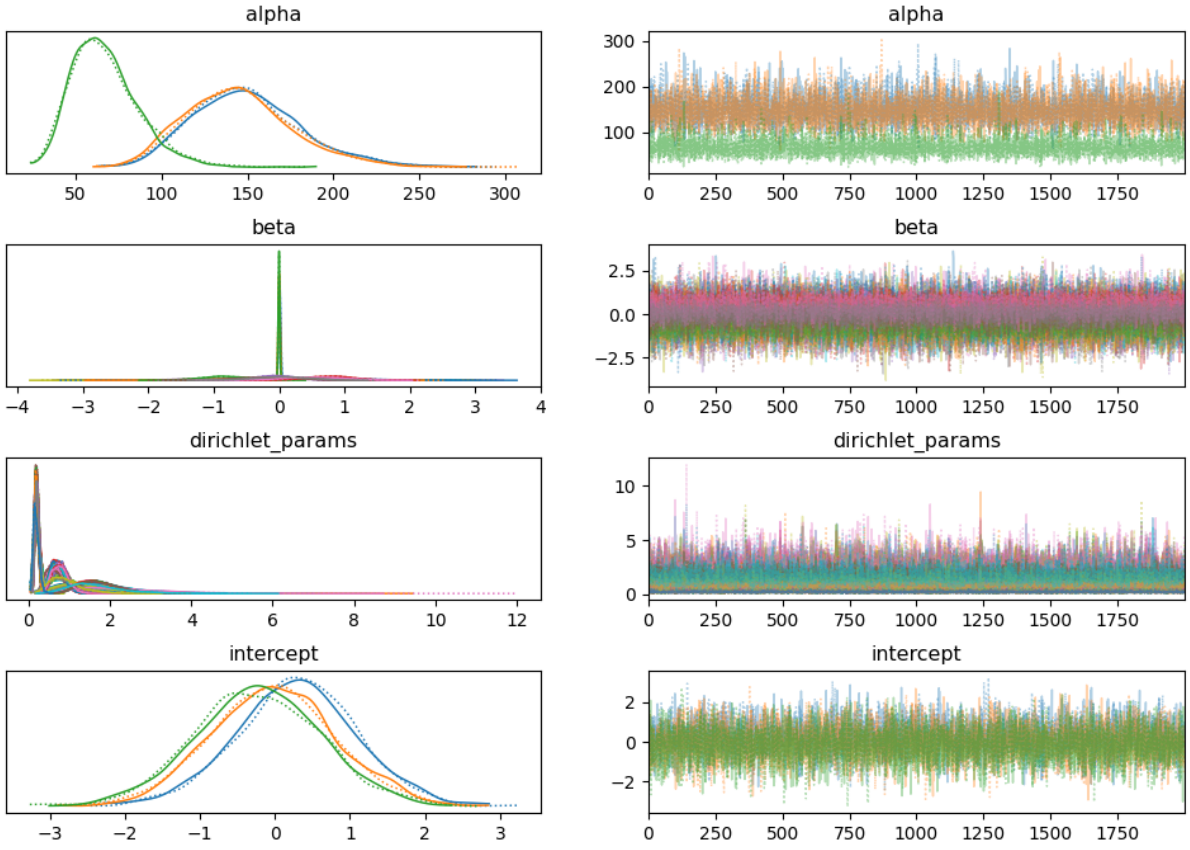

Figure 1: The Trace plots for parameters alpha and beta.

The chains for all parameters (alpha, beta, and intercept) show random fluctuations

without clear trends, indicating that the Markov chains are exploring the posterior distribution well.

3.1 The Bayesian model application on 2022 presidential elections

This study sought to fit a Bayesian model on real election data from Kenya's 2017 and 2022 presidential elections. However, the 2017 election results were nullified due to irregularities and the subsequent repeat election saw some candidates withdraw. As a result, only the 2022 presidential election results were used for this analysis. The Bayesian model was applied to investigate the impact of various socio-political factors on the election outcomes.This section presents the model estimation results, including posterior predictive checks, regression coefficient interpretations, and model diagnostics to assess the quality of the Bayesian estimation.

3.1.1 Posterior Predictive Check

To assess the fit of the model, a posterior predictive check (PPC) was conducted, comparing the simulated data from the model with the observed election results. The Bayesian p-value obtained was 0.487, which is close to 0.5. A Bayesian p-value close to 0.5 suggests that the model accurately reproduces the observed election data (Rivest and Shen (2012)). If the Bayesian p-value were too close to 0 or 1, it would indicate a poor fit of the model, implying that the simulated election results significantly deviate from reality (Rivest and Shen (2012)). Since the value obtained (0.487) is within the acceptable range, the model adequately captures the variability in voting behavior. The figure below shows posterior predictive and observed data.

3.1.2 Dirichlet Concentration Parametersα

The Dirichlet concentration parameters α represent the expected variability in vote shares for each candidate. Higher α values indicate greater stability, while lower α values suggest fluctuating voter preferences. The table below shows the concentration parameters.

Table 1: Dirichlet Concentration Parameters (α) for Presidential Candidates

Parameter	**Mean**	**SD**	**95% HDI (Low)**	**95% HDI (High)**	**ESS**	**R-hat**
α (Raila Odinga)	148.84	30.89	92.44	207.46	6282	1.0
α (William Ruto)	145.72	31.85	91.14	208.76	6105	1.0
α (Other Candidates)	67.04	19.37	35.51	103.79	7550	1.0

Note. ESS = Effective Sample Size; R-hat = Convergence Diagnostic.

From the table above, Raila Odinga ($\alpha = 148.84$) and William Ruto (α= 145.72) have high values of α, meaning that their share of votes was stable across the regions. Other Candidates (α= 67.04) have a much lower concentration value, suggesting that their support base fluctuated significantly. R-hat values = 1.0 confirm model convergence.

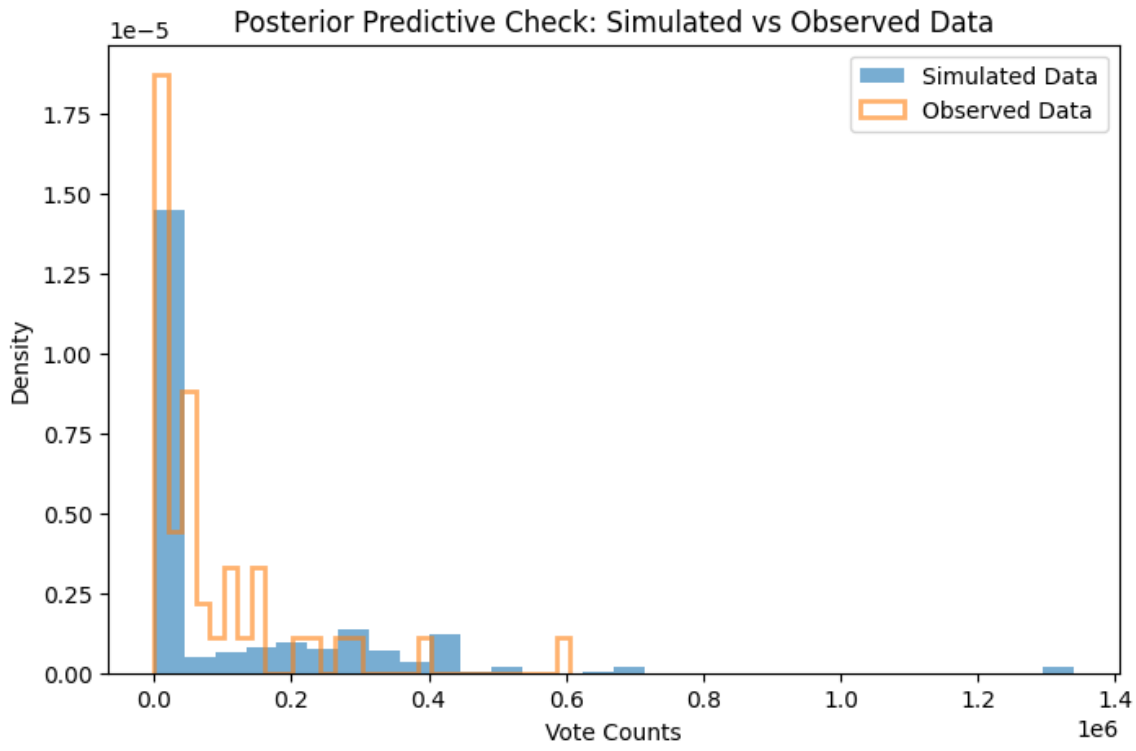

Figure 2: The model simulated and observed data.

3.1.3 Socio-political factors on the vote share among candidates

The table of regression coefficients (β) table below represents the effect of different socio-political factors on the vote shares of Raila Odinga, William Ruto, and Other Candidates.

Table 2: Regression Coefficients for Presidential Candidates

Predictor	**Raila Odinga** (β_0)	**William Ruto** (β_1)	**Other Candidates** (β_2)
Poverty Rate (%)	0.008 (0.014)	-0.001 (0.013)	-0.003 (0.012)
Voter Turnout (%)	-0.104 (0.873)	-0.424 (0.847)	-0.336 (0.859)
Female (%)	0.193 (0.953)	-0.173 (0.983)	-0.088 (0.947)
Youth (%)	-0.106 (0.847)	**0.348* (0.875)**	-0.199 (0.820)
Political Ideation (KK)	**-0.823* (0.370)**	**0.753* (0.354)**	-0.068 (0.314)
Political Ideation (Unflipped)	-0.205 (0.479)	**0.509* (0.489)**	0.000 (0.444)

Table 3: *

Note. Standard deviations are in parentheses. *Significant at 95% HDI (excludes zero).

From the coefficients table above, Youth Population (%) significantly favored William Ruto ($\beta_1 = 0.348$), confirming his stronger appeal among younger voters. Political Ideation (Kenya Kwanza) had the largest effect on William Ruto's vote share ($\beta_1 = 0.753$), showing strong partisan alignment. Political Ideation (Unflipped Voters) had a moderate positive effect on William Ruto ($\beta_1 = 0.509$), indicating that undecided voters

tended to favor him. poverty, voter turnout and gender were observed to be insignificant in influencing the voting behavior.

3.2 Bayesian model prediction accuracy against the 2022 elections opinion polls prediction

This study sought to compare Bayesian model prediction accuracy against the 2022 elections opinion polls prediction. The table below represents a comparative analysis of three election forecasting approaches—Bayesian model predictions, pre-election opinion polls (Ipsos and TIFA), and actual election results from the 2022 Kenyan presidential election. Bayesian model demonstrated higher predictive accuracy. The model estimated Raila Odinga's vote share at 48.17%, which closely matched the actual election result of 48.85%. It slightly underestimated William Ruto's vote share at 46.96%, compared to his actual 50.49%, while overestimating the vote share for Other Candidates at 4.87%, as opposed to the actual 0.66%. Despite this minor discrepancy, the Bayesian model still outperformed both Ipsos and TIFA polls, which significantly underestimated the leading candidates' vote shares and overestimated minor candidates' support. The result revealed that the bayesian model had the lowest weighted error at 8.42%, compared to 22.68% for Ipsos and 50.68% for TIFA.

4 CONCLUSIONS

The results of this study indicate that Bayesian inference is a highly effective method for predicting election outcomes. By incorporating both historical data and real-time electoral variables, the Bayesian model was able to generate more accurate predictions compared to traditional polling methods.This study shows that political ideation remains the strongest determinant of voting behavior, reinforcing the idea that voter alignments in Kenya are largely shaped by political party affiliations and coalition structures. The youth population emerged as a significant predictor of election outcomes, with the model indicating that William Ruto benefited the most from younger voters.

The comparison between Bayesian modeling and opinion polls highlights the limitations of traditional polling methodologies, particularly their tendency to overestimate support for minor candidates and underestimate last-minute shifts in voter preferences. Unlike opinion polls, which rely on fixed sample sizes and may suffer from sampling bias, Bayesian models continuously update predictions based on newly available data, making them more adaptable to real-world electoral changes.Bayesian modeling has proven to be a reliable and effective electoral forecasting tool, with the potential to revolutionize how election predictions are conducted in Kenya and beyond.

COMPETING INTERESTS DISCLAIMER:
Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

References

Alexander, B. and Ellingson, L. (2019). Poll-based conjugate prior models for the prediction united states presidential elections. In *JSM Proceedings*, pages 112–131.

Anuta, D., Churchin, J., and Luo, J. (2017). Election bias: Comparing polls and twitter in the 2016 us election. *arXiv preprint arXiv:1701.06232*.

Brass, P. R. (2023). *Ethnic groups and the state*. Taylor & Francis.

Bratton, M. and Kimenyi, M. S. (2008). Voting in kenya: Putting ethnicity in perspective. *Journal of Eastern African Studies*, 2(2):272–289.

Cameron, L. and Crosby, M. (2000). It's the economy stupid: Macroeconomics and federal elections in australia. *Economic Record*, 76(235):354–364.

Campbell, J. E., Norpoth, H., Abramowitz, A. I., Lewis-Beck, M. S., Tien, C., Erikson, R. S., Wlezien, C., Lockerbie, B., Holbrook, T. M., Jerôme, B., et al. (2017). A recap of the 2016 election forecasts. *PS: Political Science & Politics*, 50(2):331–338.

Costa, P., Nogueira, A. R., and Gama, J. (2021). Modelling voting behaviour during a general election campaign using dynamic bayesian networks. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 524–536. Springer.

Ferree, K. E., Gibson, C. C., and Long, J. D. (2014). Voting behavior and electoral irregularities in kenya's 2013 election. *Journal of Eastern African Studies*, 8(1):153–172.

Fossey, E., Harvey, C., McDermott, F., and Davidson, L. (2002). Understanding and evaluating qualitative research. *Australian New Zealand journal of psychiatry*, 36(6):717–732.

Gherghina, S. and Nemčok, M. (2021). Political parties, state resources and electoral clientelism.

Graefe, A., Küchenhoff, H., Stierle, V., and Riedl, B. (2015). Limitations of ensemble bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3):943–951.

Greenop-Roberts, H. (2022). Forecasting federal elections: New data from 2010–2019 and a discussion of alternative and emerging methods. *Australian Economic Review*, 55(1):25–39.

Grover, P., Kar, A. K., Dwivedi, Y. K., and Janssen, M. (2019). Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145:438–460.

Hassan, M. (2017). The strategic shuffle: Ethnic geography, the internal security apparatus, and elections in kenya. *American Journal of Political Science*, 61(2):382–395.

Hill, S. J. (2017). Changing votes or changing voters? how candidates and election context swing voters and mobilize the base. *Electoral Studies*, 48:131–148.

Kaur, I., Sandhu, A. K., and Kumar, Y. (2022). Artificial intelligence techniques for predictive modeling of vector-borne diseases and its pathogens: a systematic review. *Archives of Computational Methods in Engineering*, 29(6):3741–3771.

Kiingati, J. K. (2022). *A Bayesian Model for Forecasting the Choice of Candidate in a Presidential Election*. PhD thesis, JKUAT-COPAS.

Lewis-Beck, M. S. and Dassonneville, R. (2015). Forecasting elections in europe: Synthetic models. *Research & Politics*, 2(1):2053168014565128.

Montalvo, J. G., Papaspiliopoulos, O., and Stumpf-Fétizon, T. (2019). Bayesian forecasting of electoral outcomes with new parties' competition. *European Journal of Political Economy*, 59:52–70.

Noor, A. H. (2020). *Determinants of citizens' trust levels in election management bodies: a study of the Independent Electoral and Boundaries Commission (IEBC)*. PhD thesis, Strathmore University.

Paul, W., Ndati, N., Faith, M., and Siringi, S. (2016). The role of communication in creating awareness about electoral opinion polls in kenya.

Rigdon, S. E., Sauppe, J. J., and Jacobson, S. H. (2015). Forecasting the 2012 and 2014 elections using bayesian prediction and optimization. *SAGE Open*, 5(2):2158244015579724.

Rivest, R. L. and Shen, E. (2012). A bayesian method for auditing elections. In *EVT/WOTE*.

Shirima, M. A. (2018). Politics of choice: Interrogating the place of ethical decision making in kenyan politics.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.