Structural and Trial-Heat Model Combinations to Forecast US Elections

Marco Zanotti
University of Milano Bicocca

July 3, 2023

Abstract

Forecasting presidential elections is a well-studied topic given its relevance within the political field. Producing timemy and accurately prediction of the final election outcome is still not an easy task. Many different methods have been developed over the years, trying to take into account different sources of data available, from fundamental economic and political variables to state-level poll data. These methods evolved from simple structural models based on historical data to more complex Bayesian approaches that do try to exploit all the available information. Although no consensus has been reached on what is the best possible solution, it is clear that by treating forecasting as a Bayesian problem, it is possible to easily incorporate new poll data, accounting for changes in public opinion and possibly improving the predictions over structural or polling models alone.

Keywords: election forecast, bayesian modelling, trial-heat, polls

1 Introduction

Election forecasting is a strategical task in the political area since political forces spend millions of dollars in candidate's campaign and need to know when and where to allocate them. The national vote has always been considered predictable, because voters are considered to base their decisions on relatively known fundamental economic and political variables. These fundamental variables measure individuals interests and include economic conditions, party identification, ideology etc, and all the serious forecasting methods try to predict the elections using some versions of these variables. Nevertheless, producing election predictions timely and accurately in the election year is empirically hard. Firstly, close elections are always hard to predict since in these cases the best possible forecast is statistically indistinguishable from 50%. Candidates may also be close on fundamental issues that voters are more likely to base their decision on minor proposals that do actually separates them. Secondly, state-level predictions are more difficult because of geopolitical factors (there exists well-known and very important "swing" states). Moreover, uneven campaigns, low-informed elections, instability of a multi-candidate races, campaign events (such as verbal slips, gaffes, debates, etc) may also affect voters' perceptions of the candidates' positions on fundamental issues. Therefore a forecasting method based solely on some fundamental variables measured before the campaign may not work well, because these so-called structural models have no mechanism for updating predictions once new information becomes available. Nowadays this type of data is actually available, even at the state-level, that is trial-heat polls by pollsters institutions. Indeed, pre-election polls provide current information that can be used to correct potential errors in historical forecasts, increasing precision and reducing uncertainty (but have to be disregarded as literal forecasts). Therefore, a more recent and useful strategy is to use polls data to update baseline forecasts produced by structural models in a Bayesian manner, exploiting the fact that prior information are actually available in election forecasting problems. Furthermore, the Bayesian approach agrees with the well-accepted "enlightened preferences" hypothesis proposed by Gelman & King (1993), stating essentially that voters base their vote on fundamental variables and the function of the campaign is to inform individuals about them and their appropriate weights, implying that public opinion changes over the election year.

This article aims at illustrating some of the most relevant approaches to forecast (US) elections that have been developed over the years. In section 2 the most widely used type of data are discussed, and in section 3 the main frameworks of structural, trial-heat and Bayesian models are presented. Finally some considerations on some ongoing improvements are introduced.

2 Data

Forecasting elections makes use of mainly two different types of data: the so-called fundamental indicators, that is economic or political variables, and polls data. The former are historical time series covering various aspects of US economy and politics, and are usually available within 6 months of the Election day. The latter are the surveys issued by official pollsters' agency that includes trail-heat questions (i.e. at least a question on vote preference between the two major parties). In recent times, all these types of data is often available in both national and state levels.

2.1 Fundamental Indicators

Numerous researchers over many decades discovered and analysed the importance of some economic variables that strongly affect and anticipate election results. In particular, economy usually matters since an in-party presidential candidate running in the context of a booming economy would win a greater share of the vote than with a sluggish economy. Among the most used economic indicator there are GDP, GNP, unemployment, inflation at national or state level. The political dimension of election is also, obviously, of high relevance and it is usually measured by incumbency, votes of previous elections, presidential home-state advantage, partisanship of a state (proportion of democrats in last legislature), president approval rating, distance between state and candidate ideologies, and the time-for-change variable (if a party has controlled the White House for two or more terms). Sometimes also regional variables have been adopted to highlight southern and northern differences. Many models have been developed using only such data and predicted the results within few percentage points.

2.2 Trial-Heat Polls

Election polls were published in the US since the 20th century. Usually, survey data before 1988 are from Gallup, then other polling organizations emerged and started to be used too as data sources. Moreover, initially polls on presidential elections were only national, nowadays instead voters are interviewed on a state basis. Literature evidence exists to conclude that survey responses are related to actual voting process, meaning that polls are connected to observable political behaviours and incorporate the process of updating information of individuals, so that can be used to track the evolution of preferences over time and states. Election polls data suffers of some well-known problems such as sampling errors (representativeness), house effect (or organization bias, i.e. different organizations produce results systematically supporting some party), question wording, response errors, non-response bias, horse-race bias and high variability (especially during the campaign and at the state level). Nevertheless, biases arising from such effects usually cancel out by averaging over multiple concurrent surveys by different pollsters, that can be safely merged to study trends in major parties support but not undecided or not responding (Gelman & King 1993). Availability, especially of state level polls, is less an issue nowadays since many pollster agencies exist, producing numerous polls results, in particular during the election year. Moreover, a non-obvious benefit in using also trial-heat polls is that this data indirectly incorporate the more recent economic changes, since voters are considered to update their preferences based on the underlying fundamental factors.

3 Methods & Models

Given the relevance of the topic, many methods have been proposed over the years addressing the issue to produce timely and accurate forecasts of election's outcomes. Usually, the variable of interest represents the percentage election outcome of one of the two major parties (Democratic or Republican), and undecided or non-major party vote are often discarded or evenly divided. The evaluation of the models is often based on Campbell (1996) accounting method in which less than 1% is "accurate".

3.1 Structural Models

Since the 80s, simple econometric models based on structural (or fundamental political and economic) variables gained success. One of the most successful was proposed by Abramowitz in 1988 (and re-proposed in 1996 and 2008). The Time-for-Change model (Abramowitz 2008) assumes that a presidential election is essentially a referendum on the performance of the incumbent party, implying that voters are strongly influenced by their evaluation of the incumbent president's performance. Moreover, the underlying hypothesis of this model is that individuals positively evaluate periodic government alternation of the two major parties.

$$Y_t = \beta_0 + \beta_1 GDP_{t-1} + \beta_2 Approval_t + \beta_3 TC_t + \epsilon_t$$

This way, the estimate of the percentage of the incumbent party's share is based on three fundamental variables only: the second quarter growth rate of GDP, the approval rating of incumbent president and length of time the incumbent president's party has controlled the White House (time for change factor). Although this model provided relatively accurate forecasts both in 6 and 2 months before the Election day, as Gelman & King (1993) pointed out, one of the problems of models based solely on economic and political indicators is that they are based on a single regression specification relying only on previous elections' data. Hence, historical models do not incorporate in any way the opinion about the actual election that, instead, would be available by using the election poll data. Moreover, also more recent economic changes are difficult to incorporate directly through economic variables since this data is usually not available and one has to rely on past values only.

3.2 Trial-Heat Models

It is well-known that using trial-heat polls as literal forecast produce very poor results, because the accuracy of election polls in forecasting the share of votes depends enormously on when, during the election year, the poll is conducted. Indeed, it is commonplace to consider early polls as useless (same as flipping a coin) and late polls as obvious (Campbell 1996), however this data may be exploited to improve predictions.

Gelman & King (1993) proposed to incorporate actual polls information within a more complex structural model considering the aggregate trial-heat two months before the election, incumbency, GNP rate, approval rating, state specific variables (the last two state's election results, home advantage, partisanship, ideology and distance between the state and the candidate ideology), and some regional variables. The novelty of this approach rely on the fact that the authors proposed a model allowing to estimate the share of votes in each state. However, polls data was used as a national information and the predictions were produced 2 months before the elections only.

Campbell (1996), instead, improved the poor trial-heat literal prediction suggesting a simple regression model that uses only trial-heat polls at national level and the second quarter growth rate of GDP, and obtaining a forecasting performance comparable to that of previous methods, but at national level only.

3.3 Bayesian Models

Since the late 90s, methods implementing a Bayesian approach have been introduced also in the context of election prediction. The main reason is that Bayesian models naturally follow the "voters' enlightenment" hypothesis because the weights voters attach to fundamental variables are allowed to change during the campaign, accounting for changes in public opinion. The core idea of the proposed Bayesian models is to use polls data to update historical forecasts, improving the performance of structural models through the incorporation of voters preferences' evolution. Moreover, Bayesian models can often be used to estimate and study also public opinion trends nationally or at a state-level.

Brown & Chappell (1999) proposed a three-equation model where allowing poll data to be assimilated in a timely manner to update an earlier historical forecast. The *hist* equation represents the historical model, in which voting outcomes are related to structural variables (they used the growth rate of GDP in the first two quarters of the election year and the incumbency dummy). The *poll* equation, instead, is the polling model, in which voting outcomes depends on the percentage of survey respondents for that party (in the general version the authors considered also the length of the interval between poll date and election day).

$$Y_t^{hist} = \beta X_t + \epsilon_t \qquad Y_t^{poll} = \alpha_0 + \alpha_1 S_t + u_t$$

The final prediction for the election outcome is a weighted average of the historical and the poll estimates, where the weights, w^{hist} and w^{poll}, are based on the proportion of the variances of the error terms of the historical and polling regressions (i.e. the expectation of the normal posterior given normality assumption of the prior).

$$Y_t = w^{hist} Y_t^{hist} + w^{poll} Y_t^{poll}$$

Through this formulation the historical forecast is constantly updated as new poll information are available. On average, from 1952-1992, this strategy outperformed the forecasts produced by structural models and literal polling alone.

However, it is also reasonable to assume that the beliefs about election's outcomes are based on historical voting trends. Following this assumption, Steven E. Rigdon (2009) recently introduced a state election model in a fully Bayesian framework, considering also the proportions of third-party candidates and undecided. The authors developed a model that uses informative prior (based on previous election results) and current likelihood (based on ongoing poll data) for each state to estimate the posterior distribution, that is each candidate's probability of winning that state. In particular, the posterior $h(p|X)$ is built such that the likelihood $l(X|p)$ dominates the prior $f(p)$ because, as the election day approaches, poll data is more reliable than historical trends. In their formulation, being p_i the shares in a state of candidate i, the random vector of sample proportions in a state poll for n respondents is distributed as a Multinomial.

$$X = (X_1, X_2, X_3, X_4) \sim MULTINOMIAL(n, p_1, p_2, p_3, p_4)$$

Moreover, the proportions p_i are assumed to be continuous in $[0, 1]$, to satisfy $\sum_{i=1}^{4} p_i = 1$ and their joint distribution has to be a conjugate prior for a Multinomial. Hence, p is assumed to follow a Dirichlet

$$p = (p_1, p_2, p_3, p_4) \sim DIRICHLET(b_1, b_2, b_3, b_4)$$

$$p_i \sim BETA(b_i, \sum_{k=1}^{4} b_k - b_i)$$

and each p_i is distributed as a Beta random variable. Using Bayes' theorem it is possible to derive the posterior distribution, which by conjugacy is again a Dirichlet with updated parameters.

$$h(p|X) \sim C \cdot f(p) \cdot l(X|p)$$

$$h(p|X) \sim DIRICHLET(x_1 + b_1, x_2 + b_2, x_3 + b_3, x_4 + b_4)$$

The calibration and the choice of parameters are based on historical election reasoning. For instance, normal votes (i.e. votes from last elections) are used for the two major parties, while for third-party is the combined third-party normal vote, and the level of undecided is assumed to be 3% by previous polls' trends. The major contribution of this model is the incorporation of the uncertainty given by third-party preferences and undecided, while the major drawback is the absence of structural variables.

(Lock & Gelman 2010) followed a similar approach to estimate a posterior distribution for the Democratic vote share in each state, combining then the estimates to obtain the national results. The authors assumed normality of both the prior (based on historical election results), justified by the general lack of outliers in state election results, and the likelihood (based on the poll data), justified by the large sample size of each poll.

$$d_{i,0}|d_{i,t-1} \sim N(d_{i,t-1}, \sigma^2_{d_{i,0}|d_{i,t-1}})$$

$$d_{i,t}|d_{i,0} \sim N(d_{i,0}, \frac{p_{i,0}(1-p_{i,0})}{n_{i,t}} \sigma^2_{d_{i,t}|d_{i,0}})$$

The prior gives a distribution for the state i share of vote in the current election given each state's share of vote in the previous election, while the likelihood gives the distribution of a state poll, conducted t months before the election given the state's share of vote in the current election. Also in this case, the parameters are estimated using historical election results (for the prior) and historical poll data (for the likelihood). The posterior distribution is then obtained by combining the prior with the likelihood, giving a normal-normal mixture model which allows to continuously update each state's share of vote as new polls are available. Although this approach does not consider directly any fundamental variable, it produced very accurate results in forecasting 2008 US election.

(Linzer 2013) combined several aspects of previous methods in a more complex Bayesian model. The quantity of interest is still the Democratic share of vote in each state and it is estimated unifying historical forecasts based on structural variables (as opposed to use past election results only) with state-level poll data. The model has a backward component, which aggregates polls to derive estimate voting preferences, and a forward component, which predicts how these voting preferences evolve over time. In each state, voting preferences follow a reverse random walk: starting on election day, the estimated intentions evolve randomly going back in time. For the k survey, the number of respondents for the Democratic party is drawn from a Binomial distribution

$$Democratic\ Votes^k \sim Binomial(n^k, \pi_{ij}^k)$$

with poll sample size n^k and proportion of voters in a certain state i on specific day j (for $j = 1, ..., J$) π_{ij}. Hence, the quantity of interest is the share of voting preferences which is modeled as

$$\pi_{ij} = logit^{-1}(\beta_{ij} + \delta_j)$$

β_{ij} represents the historical voting preferences in state i, while δ_j is a national effect capturing variations in β_{ij} due to campaign and other election events. This model simply decomposes the log-odds of Democratic shares into a national and a state-specific component. Moreover, δ_j is set to zero on the day of the last available poll j_{last}, implying that predicting share of votes from after that day and until the election day is based on β_{ij} only. In order to smooth voting preferences by state and obtain an estimate for those days when no polls are available, two reverse random walk normal priors for β and δ are used

$$\beta_{i,j-1} \sim N(\beta_{i,j}, s_\beta^2) \qquad \delta_{j-1} \sim N(\delta_j, s_\delta^2)$$

with estimated variances that measures the rates of daily change and follow Uniform priors. The historical forecasts h_i, instead, are obtained using the Abramowitz (2008) Time-for-Change model and are incorporated through an Normal prior for the final outcome β_{iJ}

$$\beta_{iJ} \sim N(logit(h_i), \sigma_i^2)$$

where $\tau_i = 1/\sigma_i^2$ is the prior precision indicating the level of certainty on the historical forecasts, that is higher τ_i, higher weight to historical prediction over new polling data. Thus, from election day to the day of the latest available poll j_{last}, voting preferences by state start at βiJ (since $\delta_J = 0$) and follow a reverse random walk. Therefore, the Election day forecast for each state is a combination of the most recent polls and the structural forecasts and the posterior probability the Democratic candidate wins in state i is calculated as the proportion of draws from π_{iJ} greater that 0.5. In particular, when the election is soon estimates of π_{iJ} are based primarily on poll data, whereas if the election is farther ahead are driven by the historical prior (and if the precision is high, then β_{ij} converges to $logit(h_i)$ and π_{iJ} converges to h_i). Using this model, starting from final 6 months of the election year (for availability of fundamental variables), it is possible to continuously update the forecasts of the final election as new polls data is available, improving the performance over both the baseline structural model and the literal poll predictions (at least in 2008 elections). However, forecast errors are larger in infrequently polled states and it is not taken into account the uncertainty produced by third-party and undecided voters (especially important in close elections).

4 Conclusion

By treating forecasting as a Bayesian problem, it is possible to produce continuously revised forecasts as new poll data is released during the election year. This allows to account for changes in public opinion, updating the weights voters assign to the fundamental variables and possibly improving the predictions. However, comparing the performances of the different models is difficult because they are often tested against different elections, over different horizons, with different purposes (national or state levels). Moreover, more recent approaches cannot be tested in past scenarios since state-level polls were very infrequent. Uncertainty in the forecasts need also to be taken into account to effectively compare the performances of all the models.

Table 1: Reported forecasting errors over different elections.

Election	Abram.	Gelman	Campbell	Brown	Rigdon	Lock	Linzer
1988	-	-	0.7%	-	-	-	-
1992	0,9%	0.3%	0.1%	-	-	-	-
1996	0.0%	-	-	0.4%	-	-	-
2000	1.1%	-	-	-	-	-	-
2004	2.5%	-	-	-	-	-	-
2008	0.6%	-	-	-	-	0.1%	0.3%

In general, forecasts are accurate within 2 months before the election day and forecasting using both structural variables and poll data usually outperform those based on fundamentals or polls alone, both at the national and the state levels. The best improvements in using polls data are from 1 to 2 months before the election which is not really an early forecast. This is mainly due to the fact that early polls have very little relationship with the final outcome, meaning that it is still difficult to produce timely and accurate forecasts relying on this type of data. Furthermore, problems arise in forecasting accuracy and uncertainty for states that are polled few and in those days with no polls at all. To solve this lack of data issues, it is possible to use another source of spontaneous public opinions, that is web data (Rodrigue Rizk 2023). Social networks, blogs and forums contain a huge amount of data related to individuals' preferences that can be exploited during the election period to estimate the share of vote for all the candidates over all the states (thanks to geo-location feature of this data) during the election year. Finally, none of these methods consider the fact that vote intentions may be correlated across states. For example, if a candidate is performing bad in a state, this might indicate that he will also underperform in other (swing) states (possibly because there were hidden voters for the opposing candidate that polls missed in that state and may equally have been missed in others). This form of state dependency may exist and can be used to improve early accuracy of the model.

References

Abramowitz, A. I. (2008), 'Forecasting the 2008 Presidential Election with the Time-for-Change Model.', *PS: Political Science and Politics* **41**(4), 691–695.

Brown, L. B. & Chappell, H. W. J. (1999), 'Forecasting presidential elections using history and polls.', *International Journal of Forecasting* **15**(2), 127–135.

Campbell, J. E. (1996), 'Polls and Votes: The Trial-Heat Presidential Election Forecasting Model, Certainty, and Political Campaigns', *American Politics Research* **24**(4), 408–433.

Gelman, A. & King, G. (1993), 'Why Are American Presidential Election Campaign Polls So Variable When Votes are So Predictable?', *British Journal of Political Science* **23**(1), 409–451.

Linzer, D. A. (2013), 'Dynamic Bayesian Forecasting of Presidential Elections in the States', *Journal of the American Statistical Association* **108**(501), 124–134.

Lock, K. & Gelman, A. (2010), 'Bayesian Combination of State Polls and Election Forecasts.', *Political Analysis* **18**(3), 337–348.

Rodrigue Rizk, e. a. (2023), '280 Characters to the White House: Predicting 2020 U.S. Presidential Elections from Twitter Data.', *Comput Math Organ Theory* .

Steven E. Rigdon, e. a. (2009), 'A Bayesian Prediction Model for the U.S. Presidential Election.', *American Politics Research* **37**(4), 700–724.