Dynamic Forecasting of US Elections

Marco Zanotti
University of Milano Bicocca

June 25, 2023

Abstract

The text of your abstract 200 or fewer words.

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

1 Introduction

Purpose Background Challenges Gaps Directions

point out any controversy in the field

Voters at least base their decisions on relatively known and measurable variables [gelman 1993] These fundamental variables measure their interests and include economic conditions, party identification, proximity of the voter's ideology and issue preferences to those of the candidates, etc. All the serious forecasting methods try to predict the election result using some versions of the same fundamental variables to measure economic well-being, party identification, candidate quality and so forth.

1.1 Elections may be hard to predict

Nonostante la previsione nazionale è considerata essere prevedibile per via del fatto che il risultato è considerato essere basato su variabili fondamentali che sono in place before the election campaign (for instance the economic situation of the US and forte senso di appartenenza dei cittadini americani ad uno dei due partiti)

First, close elections will always be hard to predict since in these cases the best possible forecast will be statistically indistinguishable from 50%.

In primaries, low-visibility elections, and uneven campaigns, or uninformed elections we would not expect forecasting based on fundamental variables meas- ured before the campaign to work. The fast-paced events during a primary campaign (such as verbal slips, gaffes, debates, particularly good photo, opportunities, rhetorical victories, specific policy proposals, previous primary results, etc) can make an important difference because the can affect voters' perceptions of the candidates' positions on fundamental issues. Also, primary

election candidates often stand so close on fundamental issues that voters are more likely to base their decision on the minor issues that do separates the candidates.

Moreover, the inherent instability of a multi-candidate race.

Difficulty within some well-known states.

The outcome of elections with uneven campaigns would also be hard to predict based on fundamental variables alone.

However, in the general election campaign for the president (high information, balanced campaigns) these events are ephemeral having little effect on the final outcome. gelman 1993

1.2 "Mental" Process of Voters (gelman 1993)

A well-accepted hypothesis of voters process during election is the so called enlightened preferences of **?**. Essentially, voters based their preferences on fundamental variables and the function of the electoral campaign is to inform individuals about them and their appropriate weights. Hence, individuals are not rational but use increasing amounts of information over the campaign. At the beginning of the campaign voters have low level of information and this is reflected in polls answers, while the day before the election the voters have full information. Essentially the voters information set improves over the course of the campaign.

Based on this assumption a model aiming at forecasting the presidential election correctly has to incorporate the process of "voters' enlightment", implying that, since the values of the fundamental variables do not change, the weigths respondents attach to these variables have to change during the campaign, accounting for changes in public opinion.

1.3 Others

presidential election decided nowadays in swing states so it makes sense to look at state data. Now it is possible with state polls

many state are easy other diffucult next level of sophistication is to study trends in public opinion, so bayesian

state-level pre-election poll/survey data are a new source of information for both forecasting and tracking evolution of voter preferences during campaign

Existing historical models are designed to predict presidential candidates popular vote shares at a single point in time before the election (usually 2-3 months before) using structural fundamental variables such as… Prediction from these models are subject to a large amount of uncertainty, although usually accurate at the national level. Moreover, in the event that an early forecast is in error, these models have no mechanism for updating predictions once new information becomes available.

Another problem is due to the fact that regression estimates are highly uncertain due to very small samples (few past obervations on elections results becaues of one election every 4 year)

Pre-election polls provide contextual information that can be used to correct potential errors in historical forecasts increasing precision and reducing uncertainty.

But if used as literal forecasts, those from polls are very poor.

A more recent and useful strategy is to use polls data to update historical forecasts in a bayesian manner (most of the time not in real time)

IMPORTANCE OF FORECASTING: for media to explain campaign trends to the public politacal strategies to allocated mln of dollars in campaigns of candidates

2 Data

Forecasting elections makes use of mainly two different types 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 it 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.

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. It is commonplace to consider early polls as useless (same as flipping a coin) and late polls as obvious (Campbell 1996).

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.

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).

$$Y_t^{hist} = \beta X_t + \epsilon_t$$

$$Y_t^{poll} = \alpha_0 + \alpha_1 S_t + u_t$$

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 is again a Dirichlet with updated parameters by conjugacy.

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

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

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 shares in each state, combining then the estimates to obtain the national results. The authors assumed normality of 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.

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

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

The prior gives a distribution for the state 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.

3.4 Linzer 2013

(Linzer 2013) combined several aspects of previous methods in a more comprehensive 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.

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

$$\beta_{ij}|\beta_{i,j+1} \sim N(\beta_{i,j+1}, \sigma_\beta^2)$$

$$\delta_j|\delta_{j+1} \sim N(\delta_{j+1}, \sigma_\delta^2)$$

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

$$\tau_i = 1/s_i^2$$

$$\pi_{iJ} = logit^{-1}(\beta_{iJ}) since \delta_J = 0$$

The posterior probability the Democratic candidate wins in state i is calculated as the proportion of draws from π_{iJ} greater that 0.5.

Using this model, starting from final 6 months (for availability of fundamental variables), it is possible to continuously update 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 there it does not take into account uncertainty produced by third-party and undecided (especially important in close elections).

4 Conclusion

(Rodrigue Rizk 2023)

summarize major points point out significance of results questions that still remain to address

web data + ensembling

By treating forecasting as a Bayesian updating problem, we are able to produce continuously revised forecasts as new poll data are released in the course of the campaign. Allowing to account for the process of voters and incorporating the changing weights assigned to the fundamental variables.

Forecasting using both historical fundamental variables and poll data outperform those based on fundamentals or polls alone (even at the state level)

Forecasts are usually consistently accurate in the 2 months before the election.

best improvements of bayesian approaches is from 1 to 2 months before election day so still too late

problems in states polled few and in days with no polls the approach i to use polls from other states averaging but this can cause some bias in each state estimate day by day so especially for estimating trend preferences.

Web Conversations

Table 1: Reported forecasting errors over different elections.

Election	Abramowitz	three	four	five
1988	3.45	5.00	1.21	3.41
1992	3.45	5.00	1.21	3.41
1996	3.45	5.00	1.21	3.42
2000	3.45	5.00	1.21	3.43
2004	3.45	5.00	1.21	3.43
2008	3.45	5.00	1.21	3.43
2012	3.45	5.00	1.21	3.43

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.