Dynamic Forecasting of US Elections

Marco Zanotti
University of Milano Bicocca

June 22, 2023

Abstract

The text of your abstract 200 or fewer words.

Keywords: election forecast, bayesian modelling, polls, web data

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 variables, that is economic or political indicators, 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 this data is often available in both national and state levels.

2.1 Economic and Political 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.

This data is typically available before the start of the campaign in the form of time series indicators.

2.2 Trial-Heat Polls

horse-race aspects, interpreting each short-term change in the public opinion polls as a serious change in the likely fortunes of candidates

Data before 1988 are usually from Gallup, then other polling organizations emerged and are used too.

Initially only national polls then also state level polls

One can safely merge data from the different polling organizations in order to study trends in candidate support but not the percentage of undecided or not responding. Gelman 1993

The polls converge to a point near the actual election outcome shortly before the election day

Even if early polls in most election years appear to have very little to do with the eventual outcome of the general election, much evidence exists to conclude that survey responses are related to actual voting process, notably the predictive accuracy of polls immediately after the election. Hence polls are connected to observable political behaviours and incorporates the process of updating information of individuals.

Moreover, can be used to track the evolution of preferences over time and states.

PROBLEMS: - random sampling errors (representativeness) - response errors - question wording - different organization produce systematically different results (organization bias) (house effect) - high variability in the support for the Democratic and Republican candidates - non-response bias, when the candidate is going bad selectively decide not to answer or saying they do not vote - are affected in the events of the campaign - data limitation (availability) for states but no more a big issue

ONE PRO: - indirectly incorporate more recent economic changes

now wide accessibility of data

The bias arising from such effects usually cancels out by averaging over multiple concurrent surveys by different pollsters.

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 Y_t 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 between the two major parties.

The evaluation of the models is often based on Campbell (1996) accounting method in which less than 1% is "accurate", between 1 and 2 points is "quite accurate", between 2 and 3 points is "reasonably accurate", between 3 and 4 points is "fairly accurate", between 4 and 5 points is "inaccurate", and in excess of 5 points is "very inaccurate".

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.

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

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 of all the limitations of the polls. Indeed, 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).

In 1993, Gelman and King 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 (Gelman & King 1993). 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.

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.

3.3 Bayesian Models

3.4 Brown Chappel 1999

Bayesian model that uses both polls and historical data and allows poll data to be assimilated in an optimal and timely manner to update an earlier forecast Uses time series data over elections

only simple model and generalized with description of regressors

Found gains in using polls to augment historical regression data. Weighted average forecasts have lower MSE than historical or trial-heat only.

3.5 Rigdon 2009

Presidential vote forecasting models have been heavily focused on the national two-party popular vote because of political reasons and for data availability issues. Since now state-by-state polls are available should change attention from popular vote and electoral votes, also including third-party candidates and undecided in the model.

Bayesian model that uses prior and current information for each state to determine each candidate's probability of winning that state. Uses previous election's results to capture each state's party tendency and current polling data for current tendency.

Informative prior based on long-term voting trends within a state and different probability distributions.

Without knowledge, it is reasonable that the belief about the election's outcome is based on historical trends, hence state's partisan tendency is used as informative prior. This is coupled with polling data to estimate the posterior distribution (prob of winning a state).

Updates the historical beleifs based on ongoing data (polls). Posterior is built such that likelihood dominates prior because the polls data are more reliable than historical trends to forecast elections.

h(p|X) = posterior Cb proportionality constant f(p) = prior density, Dirichelt g(X|p) = likelihood density, multinomial

prior is conjugate and posterior is still Dirchlet with updated parameter (for each state)

The choice of parameter is based on Normal Votes (i.e. votes from last elections), for third-party is the combined third party vote in last election, undecided 3% by history

use posterior to estimate pi and then uses the formula to compute electoral votes in each state

3.6 Lock Gelman 2010

Bayesian model to integrate various sources of data to form posterior distributions for states and then national two-party Democratic vote shares.
Can be used also to study changes in public opinion and spatial relationships among states.

uses historical election results data by states and historical polls to estimate appropriate weighting to use in combining surveys and produce forecasts

do not consider third-party or undecided

model more robust to pre-election fluctuations

combines estimate not specific to a certain point in time with ongoing polls, updating continuously the forecasts of public opinion.

actually a model on relative state position (difference between proportion of voting Demo-

cratic in each state and the national proportion of voting Democratic)

adjustment for home-state advantage

prior, likelihood and posterior are normal distributions

3.7 Linzer 2013

Dynamic Bayesian model that unifies regression-based historical forecasting approaches developed in political and economic sciences (based on fundamentals) with the poll tracking capabilities made feasible by the recent upsurge in state-level opinion polling.

Since the day of election comes by, then polls become more and more reliable and relevant in the predictions. Older polls are useful however to estimate public opinion trends and variations over the campaign.

Include campaign effects

quantity of interest is the democratic share of the state-level major-party vote does not include third-party or undecided

estimated swing states (those whose forecasts are very close)

from final 6 months, moving forward and updating historical initial forecasts every 2 weeks with polls data

largest forecast errors in infrequently polled safe states

4 Conclusion

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.

4.1 Web Conversations

5 Notes

(Abramowitz 2008) (Gelman & King 1993) (Campbell 1996) (Brown & Chappell 1999) (Steven E. Rigdon 2009) (Lock & Gelman 2010) (Linzer 2013) (Rodrigue Rizk 2023)

[Consistency comparison in fitting surrogate model in the tidal power example.]{#fig-first width=3in}

Table 1: D-optimality values for design X under five different scenarios.

one	two	three	four	five
1.23	3.45	5.00	1.21	3.41
1.23	3.45	5.00	1.21	3.42
1.23	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.