Predicting Elections using Live Data at The Washington Post

Leonard Bronner
lenny.bronner@washpost.com
The Washington Post

Al Johri
al.johri@gmail.com
The Washington Post

Jeremy Bowers
jeremy.bowers@washpost.com
The Washington Post

ABSTRACT

In this paper we present The Washington Post's method for forecasting turnout during an election using partial precinct results. It allows computation of pre-night predictions, which are used as a starting point for the live updating model. The method includes precinct covariance computation, which are then used to update the predictions throughout the night and for computing prediction intervals to measure the confidence of the predictions. It was designed for the general election in 2020, and was tested on the 2019 Virginia House of Delegates and state Senate election.

KEYWORDS

elections, forecasting, live model

ACM Reference Format:
Leonard Bronner, Al Johri, and Jeremy Bowers. 2019. Predicting Elections using Live Data at The Washington Post. In *Computation + Journalism Symposium 2020.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

On most election nights, the best way to communicate how many votes are outstanding is the fraction of precincts that have reported their results. Unfortunately, this metric can be misleading. For example, if many small precincts have already reported but a few large precincts haven't, the percentage of precincts reporting would give the idea that most of the election was already over and was possibly already decided. Conversely, if only a few large precincts have finished counting, it could appear that there were many votes outstanding even if election night was practically nearly over. This also doesn't take into account things like early, absentee or provisional voting, which can make up much more than half of the total vote in some areas [14].

To avoid this issue some news organizations rely on live models that try to estimate turnout while the election is still ongoing. Starting with the 1952 election, computers and algorithms have played an important part in giving readers and viewers more accurate information about the state of the race. Since then, television networks and news agencies have used these models to call elections before all the votes have been counted. More recently newspapers, such as The New York Times, and online media organizations, such as FiveThirtyEight, have begun using models on their own to give their readers a better understanding of election night.

For the Virginia elections in 2019, The Washington Post ran its first live election model. This paper will describe the model and discuss how it performed in the election.

2 RELATED WORK

Since the 2008 and 2012 elections, in which FiveThirtyEight accurately predicted Barack Obama's election and re-election as President of the United States [18] [19], election models have become a popular feature in the media landscape. 2016 saw a further proliferation of them, with The New York Times [9] and HuffPost [7] being two media organizations, among others, to release models. There were also a number of private individuals, such as Pierre-Antoine Kremp [11] and G. Elliott Morris [13], who released public election models.

There has also been academic work on the subject. Some models use only fundamental data such as economic indicators, presidential approval ratings and incumbency [6] [1] [2]. While others rely purely on polling and poll aggregation [21]. Finally, some approaches try to blend both fundamental and polling data [12].

There has been less public work on live models, even though these models have been commonly used to predict and call elections since the Presidential election in 1952 [3]. If an accurate pre-night model is present, it can be used to run thousands of simulated elections, which are then used to forecast outcomes based on the conditional distribution of these simulated elections [4]. If no pre-night model is present, an approach has been to learn the relationship between historical results and current results for precincts that have reported and then apply this relationship to outstanding precincts [17] [15]. [1]

More complex and experimental approaches include clustering the electorate into groups of similar voting behaviour and then extrapolating partial results to whole clusters and by extension the entire electorate [8]. Other approaches include using geo-spatial information [16] or genetic algorithms [5] to make election night forecasts.

[1] This approach is relatively common in newsrooms for live election forecasting

C+J '20, March 20–21, 2020, Boston, MA
.

3 MODEL

The Washington Post model has two components to it. We initially generate pre-night turnout estimates using Ordinary Least Squares. Once precincts start reporting, these estimates are updated using the conditional distribution of the outstanding precincts given the ones we have observed.

The model is founded on the assumption that past turnout predicts future turnout. While there are exceptions, this maxim allows for a good preliminary approximation of the results. Because the model was built to adapt to live data rather than act as a static pre-election prediction, these estimates are a starting point only.

The focus on the live experience is also the reason the model is on a precinct level. By returning predictions for each precinct, it does not have to wait for entire districts or counties to finish counting and report their results before updating its predictions.

Prenight Model

To learn the relationship between past turnout and future turnout, we looked at historical elections that most closely resembled this election. For the House of Delegates, we used the 2017 election, which included a Governor's race but no state Senate elections, and 2015, which on the surface is the most similar to 2019. For the state Senate, we chose the 2015 and 2011 elections.

It's possible that choosing the 2017 election was a mistake since, in a lot of important ways, it was quite different than the election this year. But we made a conscious choice to include it because we thought that an election in which Democrats nearly doubled their Gubernatorial win margin compared to 2013 [10] may have marked a fundamental shift in Virginia politics [20].

To predict the turnout for any of those elections, we looked at the turnout in prior elections. So to predict the 2017 turnout, we looked at turnout in 2016, 2015 and 2013. We also included turnout from the year's primary and the turnout in the primary of the equivalent election two years previously. Since a lot can change between elections, primary turnout gave us a better indicator of enthusiasm for an election in a specific year. Other features we added included an indicator of whether or not there was a Governor's race — which helps us account for some of the increased turnout in 2017 — and precinct-level demographic features including 10-year age buckets and ethnicity.

We applied ordinary least squares (OLS) to those features to predict turnout precinct-by-precinct in 2019.

The setup is the usual one for OLS:

$$\vec{y} = X\beta + \epsilon \tag{1}$$

Let n be the number of precincts and p the number of covariates. $\vec{y} \in \mathbb{R}^n$ is the vector of precinct turnouts and $X \in \mathbb{R}^{n \times p}$ is the matrix of covariates. $\beta \in \mathbb{R}^p$ is the vector of parameters we are trying to estimate and $\epsilon \in \mathbb{R}^n$ is the error. We assume that $\epsilon \sim \mathcal{N}(0, \Sigma)$ so by linearity of expectation $\vec{y} \sim \mathcal{N}(X\beta, \Sigma)$.

The solution to equation (1) was obtained using the Moore-Penrose pseudo-inverse:

$$\beta = (X^T X)^{-1} X^T \vec{y} \tag{2}$$

The fitted model gives us parameter estimates for the covariates specified above, which we then apply to the corresponding data in 2019. The covariates for 2019 were now the 2018, 2017 and 2015 general election turnouts, 2019 and 2017 primary turnout, an indicator of a Governor's election (set to 0) and the precinct-level demographic features — 10-year age buckets and ethnicity.

Prediction Intervals

To give the predictions an upper and lower bound, we compute prediction intervals. Prediction intervals are similar to confidence intervals, which contain the true population of a sample statistic with some predefined probability. The difference is that prediction intervals take into account the additional uncertainty of predicting future observations.

Since our predictions are the output of a linear regression it would seem reasonable to use a t-distribution for our prediction intervals. However, using this distribution would build on the assumption that the precincts are independent of each other. This is likely incorrect. The factors that drive high turnout in one precinct would, at a minimum, likely also drive high turnout in other precincts within the same district. And these same factors might even suggest high turnout generally across the entire election.

To solve this issue we used the multidimensional analog of the t statistic known as Hotelling's T-squared statistic. It gave us a multidimensional prediction region for all precincts simultaneously. This statistic follows an F-distribution, so it has thicker tails which take into account more uncertainty, giving us robust prediction intervals for each precinct.

Let $\vec{x} \in \mathbb{R}^n$ such that $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$. Assume we are trying to estimate the mean with the sample mean $\bar{\vec{x}}$ and the covariance matrix with the sample covariance matrix $\boldsymbol{S}$. Consider the test statistic:

$$T^2 = n(\bar{\vec{x}} - \vec{\mu})^T \boldsymbol{S}^{-1} (\bar{\vec{x}} - \vec{\mu}) \tag{3}$$

Then by Hotelling $\frac{n-p}{(n-1)p} T^2 \sim F_{p,n-p}$. Therfore the $1 - \alpha$ confidence ellipsis are

$$\left(\bar{\vec{x}} - \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha) \frac{S}{n}}, \bar{\vec{x}} + \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha) \frac{S}{n}}\right) \tag{4}$$

Which would give us the confidence ellipsis for $\bar{\vec{x}}$. Since we are interested in the prediction interval, we add in the estimate for the sampling variation of the new point, which is the mean squared error over our training set.

To extract simultaneous confidence intervals for any combination of the precincts (for example the districts), we simply sum over the corresponding covariance elements and then add in the estimate of the sampling variation.

Live Model

The second — and more important — component of our turnout model was a live model that could use actual turnout results from precincts that had already reported to update our predictions for those that hadn't.

There are two problems that make live models complicated. The first is that high turnout in one precinct does not necessarily mean high turnout in other, different precincts. The second issue is that precincts do not report in random order. Generally, rural precincts report faster than urban precincts, which makes extrapolating from early precincts difficult. Our approach should solve the first issue, though the second one remains.

Through the use of Ordinary Least Squares regression and the Hotelling prediction intervals we had already baked in the assumption that the vector of precinct-level predictions were jointly normal. Throughout the night, we were interested in the conditional distribution of the unreported subset of those predictions given the observed predictions in the reported precincts. Thankfully, the conditional distribution of the unreported precincts is also normal and with well-known closed form solutions for the conditional mean and covariance matrix.

Let $\vec{x} = \begin{bmatrix}\vec{x_1} & \vec{x_2}\end{bmatrix}^T$ be the precinct level vector where $\vec{x_2}$ is a sub-vector of precincts we are about to observe and $\vec{x_1}$ is a sub-vector of the rest of the precincts. By our previous assumption $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ where:

$$\vec{\mu} = \begin{bmatrix}\vec{\mu_1} \\ \vec{\mu_2}\end{bmatrix}, \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

Assume now we observe the precincts in $\vec{x_2}$ such that $\vec{x_2} = \vec{a}$. The conditional distribution of $\vec{x_1}$ given $\vec{x_2}$ is now:

$$\vec{x_1} | \vec{x_2} = \vec{a} \sim \mathcal{N}(\hat{\vec{\mu}}, \hat{\Sigma}) \tag{5}$$

where

$$\hat{\vec{\mu}} = \vec{\mu_1} + \Sigma_{12}(\Sigma_{22})^{-1}(\vec{a} - \vec{\mu_2})$$
$$\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21}$$

Intuitively, the new mean is the old mean plus the difference in prediction and observed value of the observed precincts, filtered through the covariance, which is an estimate for the relationship between the observed and unobserved precincts.

Covariance Estimation

Because our model relies heavily on covariances — both for prediction intervals and for the live updates — generating the covariance matrix became an area of intense scrutiny for us. The usual approach would have been to take the sample covariance as the covariance estimate. Unfortunately, we couldn't just do that because of the intricacies of our sparse data set of historical elections.

There are approximately 2,000 precincts in Virginia. Our covariance matrix — a 2,000 by 2,000 matrix — would have two million unique entries. Unfortunately, there have been very few — at best, four — comparable elections that we can use as samples to compute the covariances. The underspecification of this problem means that the output of the covariance estimate, if done naively, was numerically unstable. Which resulted in the the updated covariances jumping to zero after very few precincts were observed. The resulting matrices were no longer positive semi-definite and close to zero. This meant they could have produced negative district level variances, something which would have broken our prediction intervals.

We tried a number of methods to mitigate this issue, including applying an inverse-Wishart prior distribution to the covariance matrix and scaling the matrix by a multiple of the identity matrix. Unfortunately none of these methods had enough of an effect on the estimand.

We settled on two different approaches — one for the House of Delegates and one for the state Senate. We are using a different approach for each chamber because there have been fewer state Senate elections using the current districts — and many of those races were previously uncontested — so our sample covariance estimates were even less stable.

The covariance matrix in the state Senate consisted of three unique numbers: The precinct variance, the covariance between two precincts in the same district and the covariance between two precincts in different districts. These are each calculated as the average over the original sample covariances.

For the House of Delegates, we assumed that the covariance between two districts was a function of the physical distance between those two districts and the Mahalanobis distance between demographic vectors. We solved this linear equation using OLS with the sample covariance as the dependent variable. As in the Senate, we continue to use the average sample variance as our variance estimate.

The covariance estimate is:

$$\hat{\sigma}_{i,j} = \alpha + \beta \cdot d(i,j) + \gamma \cdot sim(i,j) \tag{6}$$

Where $d(i,j)$ is the distance between the centroids and of precinct i and precinct j and $sim(i,j)$ is the Mahalanobis distance between the demographic features of the precincts.

We used OLS to solve the above equation for all districts. Unfortunately this gives us no guarantees on the semi definiteness of the resulting matrix. In the future we want to explore the use of semi-definite programming for this problem.

% Prec. Reporting	House	Senate
0	1,921,904 (-10.3%)	1,899,178 (-10.86%)
25	2,074,615 (-3.2%)	2,048,599 (-3.84%)
50	2,086,871 (-2.6%)	2,105,937 (-1.15%)
75	2,127,110 (-0.72%)	2,189,102 (+2.75%)
87	2,135,681 (-0.32%)	2,214,725 (-0.35%)
92	-	2,222,429 (+4.32%)
Actual	2,142,461	2,130,441

Table 1: Predicted Turnout

% Prec. Reporting	0	25	50	75	87
MAE (Prec. Lvl)	148.46	85.55	44.72	19.36	4.44
MAPE (Prec. Lvl)	35.07	28.90	12.36	4.81	2.00
MAE (Unk Prec.)	148.14	102.51	88.91	92.71	94.17
MAPE (Dist. Lvl)	16.70	8.17	4.11	1.84	0.49

Table 2: Results for the House of Delegates

% Prec. Reporting	0	25	50	75	92
MAE (Prec. Lvl)	141.18	110.44	67.59	56.24	37.73
MAPE (Prec. Lvl)	19.84	35.34	21.90	14.30	4.12
MAE (Unk Prec.)	139.71	125.70	103.99	98.56	151.11
MAPE (Dist. Lvl)	13.83	9.66	5.66	5.29	4.20

Table 3: Results for the state Senate

4 RESULTS

Model

As can be seen in table 1 our pre-night turnout predictions were 1.9 million votes for both the House of Delegates and state Senate [2]. Throughout the night these approached (and for the state Senate overshot) the actual (non-absentee) turnout numbers. In parenthesis you can see how far off the predictions were as a percent of total turnout.

To evaluate our model, we use mean absolute error (MAE), since that metric is interpretable as raw vote differences and unlike root mean squared error (RMSE) the loss scales linearly with the error. To get a sense for the scale of the error we look at mean absolute percentage error (MAPE). We look at MAE both over all districts, since this tells us how far off we were in all races, and MAE for only the outstanding precincts, since this tells us how good our predictions were in that moment in time. Since the model made predictions on the precinct level, our main evaluation should be done on the precinct level too.

% Prec. Reporting	0	25	50	75	87
Size	1107.11	843.33	382.69	128.42	25.40
Performance	0.98	0.98	0.90	0.86	0.95

Table 4: Prediction interval coverage, House of Delegates

% Prec. Reporting	0	25	50	75	92
Size	644.37	485.21	260.75	92.99	0.15
Performance	0.90	0.90	0.86	0.81	0.91

Table 5: Prediction interval coverage, state Senate

Results at 0%, 25%, 50% and nearly 100% reporting can be seen in table 2 and table 3. [3]

Over all House of Delegates precincts the MAE went from 148 votes to 4 votes. This is equivalent to MAPE changing from 35% to 2%. Over unknown precincts only the MAE also starts at 148 votes and stabilizes at around 90 votes.

For the state Senate, the MAE over all precincts goes from 141 to 37 votes, which is equivalent to a MAPE change of 19% to 4%. For unknown precincts the error converges at around 100 votes. [4]

It seems that the errors on a precinct level cancel each other out, as can be seen in the district level results. For the House of Delegates at 0% reporting MAPE is at only 17% and is at less than 2% with only 75% reporting. In the state Senate MAPE starts at 14% and at 50% reporting is at only 5%, where it remains. This is further supported by the numbers from table 1, which show the predictions for the entire state.

Prediction Interval

To evaluate the prediction intervals we can look at their average size and the fraction of results contained within the prediction interval at any given time. As can be seen in table 4 and table 5 the prediction interval coverage remained relatively high throughout the night. We were aiming for 95% coverage, while we nearly achieved on average, the fact that the coverage dips to 86% in the House of Delegates and 81% in the state Senate is concerning.

This may have been caused by the uneven distribution of precincts reporting. Precincts that came in early are generally more rural, while the later precincts are urban. Democrats, who generally cluster in these late reporting, urban precincts had a good night in the Virginia election. This means that turnout was probably especially high in these urban precincts that came in late, making it more likely that we did not achieve prediction interval coverage in those precincts. This

[2] None of these numbers contain absentee votes

[3] 86% reporting in the House of Delegates and 92% reporting in the state Senate were the most complete results available at the time. Many precincts in uncontested districts had not reported yet on the Department of Election website.

[4] Precincts for the state Senate are larger than for the House of Delegates

would explain why the prediction interval coverage recovers towards the end, once more urban precincts have reported.

Another cause may have been changes in the degrees of freedom of the f-distribution throughout the night. As fewer precincts are outstanding, the degrees of freedom of the f-distribution used to calculate the prediction intervals should decrease. It is now however, clear how quickly this should happen (if precincts were independent, then the degrees of freedom should decrease by 1 for every precinct reported, but they aren't independent, which is why we are using an f-distribution, instead of a t-distribution, in the first place). The reason we can assume this may be a cause is that the size of the prediction intervals falls very quickly, at 50% reporting the prediction interval size is only 30% of it's size at the beginning of the night, but we would expect it to only be at 50% of its pre-night size.

5 EVALUATION

Overall, this model helped us convey to The Washington Post readers the state of the race and associated uncertainty for the 2019 Virginia House of Delegates and state Senate elections.

Since performance metrics from live election forecasts are usually not distributed by news organizations, making direct comparisons is difficult. However, the raw performance, as measured via MAE and MAPE, are competitive.

The particular upside of the model is our ability to update our prediction intervals such that they are taking into account which precincts are still outstanding and how those precincts relate to one another.

However, our reliance on the covariance matrix is also a downside of this model, as it can be difficult or numerically unstable to compute. To alleviate this problem we shrink the matrix towards some constant, which makes our updates more conservative and our model slower to adapt to early results.

We can also generally observe that that the House of Delegates model performed better than the state Senate model. This is probably related to the better covariance estimation procedure. Future work should include rerunning the Virginia election for the State Senate using the same covariance estimation algorithm used for the House of Delegates.

ACKNOWLEDGMENTS

Special thanks to Drew Dimmery, John Cherian, Idrees Kahloon and Nick Diakopoulos.

REFERENCES

[1] Alan I. Abramowitz. 1988. An Improved Model for Predicting Presidential Election Outcomes. *PS: Political Science & Politics* 21, 4 (1988), 843–847.

[2] Alan I. Abramowitz. 2016. Will Time for Change Mean Time for Trump? *PS: Political Science & Politics* 49, 4 (2016), 659–660.

[3] Ira Chinoy. 2010. *Battle of the Brains: Election-Night Forecasting at the Dawn of the Computer Age.* Ph.D. Dissertation. University of Maryland, College Park, MD.

[4] Andrew Gelman and Nate Silver. 2010. What do we know at 7pm on election night? 1.

[5] Ronald Hochreiter and Christoph Waldhauser. 2014. Evolving Accuracy: A Genetic Algorithm To Improve Election Night Forecasts. *Applied Soft Computing* 34 (01 2014).

[6] Patrick Hummel and David Rothschild. 2014. Fundamental Models for Forecasting Elections at the State Level. *Electoral Studies* 35 (09 2014).

[7] Natalie Jackson and Adam Hooper. 2016. Forecast President. https://elections.huffingtonpost.com/2016/forecast/president

[8] Chris Elphinstone JM Greben and Jenny Holloway. 2005. A model for election night forecasting applied to the 2004 South African elections. *ORiON* 22 (08 2005), 89–103.

[9] Josh Katz. 2016. Who Will Be President? https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html

[10] Dan Keating and Kevin Uhrmacher. 2017. An enthusiastic, more polarized Virginia electorate gave Northam the win. https://www.washingtonpost.com/graphics/2017/local/governor-turnout-analysis/

[11] Pierre-Antoine Kremp. 2016. State and National Poll Aggregation. http://www.slate.com/features/pkremp_forecast/report.html

[12] Drew A. Linzer. 2013. Dynamic Bayesian Forecasting of Presidential Elections in the States. *J. Amer. Statist. Assoc.* 108, 501 (2013), 124–134.

[13] G. Elliott Morris. 2016. 2016 Presidential Election Forecast. https://web.archive.org/web/20161110234247/http://www.thecrosstab.com/2016%20Forecast/

[14] National Conference of State Legislatures. 2019. Absentee and Early Voting. http://www.ncsl.org/research/elections-and-campaigns/absentee-and-early-voting.aspx

[15] Jose Manuel Pavia-Miralles. 2005. Forecasts from Nonrandom Samples: The Election Night Case. *J. Amer. Statist. Assoc.* 100 (12 2005), 1113–1122.

[16] Jose Manuel Pavía, Beatriz Larraz, and Jose Marí Montero. 2008. Election Forecasts Using Spatiotemporal Models. *J. Amer. Statist. Assoc.* 103, 483 (2008), 1050–1059.

[17] Clive Payne. 2003. Election Forecasting in the UK: The BBC's Experience. *Euramerica* 33 (01 2003).

[18] Nate Silver. 2008. Today's Polls and Final Election Projection: Obama 349, McCain 189. https://fivethirtyeight.com/features/todays-polls-and-final-election/

[19] Nate Silver. 2012. Special Coverage: The 2012 Presidential Election. https://fivethirtyeight.com/features/special-coverage-the-2012-presidential-election/#not-a-big-win-for-obama-but-a-broad-one

[20] Laura Vozzella. 2019. Virginia's 'off-off-year' elections were once sleepy. And then came Trump. https://www.washingtonpost.com/local/virginia-politics/virginias-off-off-year-elections-were-once-sleepy-and-then-came-trump/2019/09/14/860570da-cff0-11e9-b29b-a528dc82154a_story.html

[21] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. 2014. Forecasting Election With Non-Representative Polls. *International Journal of Forecasting* 31 (09 2014).