Smoothing Mixed Effects Models and Election Poll Aggregation: An Application to the 2020 Democratic Primary

Lisa Wilson

June 8, 2020

Background

Accurately forecasting elections remains a compelling statistical problem, one that has come under greater scrutiny after high profile "failures" of polls to predict election outcomes in the United States and abroad. Following the 2016 U.S. presidential election, concerns about the reliability of election polling increased, as most websites and news organizations involved in election forecasting had predicted high probabilities of a Clinton win. In examining the performance of election polls themselves, an American Association for Public Opinion Research (AAPOR) report on the 2016 election found that national polls were accurate overall, particularly compared to past presidential election polling, and that while state-level polls in battleground states reflected a close race, they underestimated support for Trump in the previously Democratic battleground states of Michigan, Wisconsin, and Pennsylvania. AAPOR attributed the contrast in performance in national and state polls, at least in part, to the tendency for state-level pollsters to have lower budgets. They also concluded that there was sufficient evidence that there was a change in voter opinion in the week before the election in key states and that polls did not adequately adjust survey weights to account for the overrepresentation of college graduates in their samples (Kennedy 2018).

Dicussions about the state of polling and work on developing more accurate election prediction models in the U.S. often focus on the general presidential election as opposed to presidential primaries. Primary elections present a number of challenges not present in general elections, such as the greater number of candidates and their respective ideological positions, which also vary across election cycles, the range of election days across states, and the variation in how states conduct primaries or caucuses and how they award delegates.

With a focus on research into modeling elections in multiparty systems as well as research into improving U.S. models following 2016, a review of the basic types of election forecasting models offers several paths for predicting primary outcomes.

Fundamental or structural models use various covariates, including economic indicators, political ideology metrics, and approval ratings, in addition to historical data such as past election outcomes, to predict upcoming elections. They do not include current polling data. One often-cited example that has performed well for past U.S. presidential elections is the "bread and peace" model from Hibbs (2000), which uses only per capita income growth and total military personnel killed in action over the current presidential term. Walther (2015) argues that fundamental models are more difficult to apply in multiparty systems because of the complexities of coalition-building between multiple parties, the problem of assigning responsibility to one party over another for changes in economic variables and other common covariates, and the difficulty of incorporating new political parties in models. Similar arguments could be made about multi-candidate U.S. presidential primary elections; for example, it becomes difficult to integrate historical data into primary models as the number of viable candidates, the ideological positions they represent, and the level of name recognition varies across primary election cycles.

Poll aggregation models aim to combine the public opinion data in various polls to forecast election outcomes. Dynamic linear models (DLM), which often involve Bayesian methods, are among the more complex examples. Beginning with a DLM that hierarchically models the observed polling data for a particular candidate in a multiparty system as the true underlying trend plus noise and the underlying trend as a random walk, Walther adds a seasonal component to capture the types of seasonal changes in support that have been observed in past elections, such as the support for the incumbent party decreasing in the middle of the term before recovering at least somewhat as the election approaches or the cycle of diminished support and gradual recovery following a scandal. Michaelis (2018) presents a model similar to Walther's DLM that incorporates a random effect for the pollster and a time

dependent effect that can be modeled flexibly as some form of ARMA model. For example, in an application to German parliamentary election data, Michaelis uses an AR(1) model for the time dependent effect. Media organizations and websites such as FiveThirtyEight, the New York Times Upshot, Princeton Election Consortium, and HuffPost rely primarily on poll aggregation for their forecasts, although the FiveThirtyEight "polls-plus" model also integrates economic and historical data. This qualifies the "polls-plus" model as a hybrid or synthetic model, which incorporates both fundamental elements and poll aggregation. The dynamic Bayesian model in Linzer (2013), which combined forecasts based on historical data with state polls to forecast the 2008 U.S. presidential election, is another example.

Wright (2018) presents a new poll aggregation approach as a more accurate alternative to the models used by the prediction sites listed above, which tended to predict a low probability of Trump winning in 2016. Using a smoothing mixed effects model with the log-ratio of percent support for Clinton over Trump in each state as the response, this approach matches or outperforms the prediction sites on a number of metrics and predicts a 47.2% probability of a Trump electoral college win, whereas the highest such predicted probability from any of the sites was 29% from FiveThirtyEight. Wright uses the log-ratio as the response in part because it is less affected by the percentage of third-party supporters, which tends to be higher in polls than in election results and can lead to inaccurate predictions. Additionally, because the log-ratio and the difference (or spread) between candidate support have the same sign, predictions about which candidate will win a particular state can be made on either scale. In addition to the model's reliance only on available polling data, the reduced impact of non-major candidate preference makes this model appealing in a multi-candidate primary race where frontrunners representing different ideological positions emerge ahead of other candidates, as seen in the 2020 Democratic presidential primary. Accordingly, the performance of this smoothing mixed effects model is evaluated using polling data for states that held 2020 Democratic primary elections on Super Tuesday.

Smoothing Mixed Effects Model

The smoothing mixed effects model treats a transformed value of candidate support in a particular state as a smooth function over time that can be decomposed into an overall national effect as well as effects specific to the state. It is defined as

$$y_{ij}(t) = \mu(t) + v_i(t) + \epsilon_{ij}$$

where $y_{ij}(t) = \log(\frac{A_{ij}(t)}{B_{ij}(t)})$ is the log-ratio of percent support for candidate A to candidate B for state i and poll j at time t days before the election, $\mu(t)$ is the national fixed effect (as measured from state polls, with no additional national polls used), $v_i(t)$ is the state-level random effect, and ϵ_{ij} are iid normal error terms with mean 0. Then let $\theta_i(t) = E(y_i(t))$ denote the true log-ratio of support for candidate A to candidate B, with $\hat{\theta}_i(t)$ estimated from polling data. To obtain the estimates $\hat{\theta}_i(t)$, the model is fit using polling data and then spline extrapolation is used to find estimates for the particular time t. $\hat{\theta}_i(0)$, for example, is the spline-extrapolated estimate of the log-ratio of candidate support on election day in state i. Following Wright and other election forecasting models, the time t used when fitting the model is the number of days between the election and the mid-date of the period over which the poll took place.

Model estimation is performed with the R package `sme`, which fits the smooth functions $\mu(t)$ and $\{v_i(t)\}$ with splines using maximum likelihood estimation that is penalized by bandwidth parameters λ_μ and λ_v. To select appropriate values for these bandwidths, the available polls are split into a training and test set, with the test set comprising the final 10% of polls by date. Models are then fit using all possible combinations of $\lambda_\mu, \lambda_v = (0.5, 1, 5, 10, 50, 100, 500, 1000, 5000)$, spline extrapolated estimates for the dates in the test set are obtained based on each model, and the root mean square error (RMSE) between the estimated log-ratio and the true log-ratio is calculated for each model. The $\{\lambda_\mu, \lambda_v\}$ combination that minimizes RMSE is then used to fit a model using all the polling data,

after which spline-extrapolated estimates of the log-ratio on election day are obtained.

To calculate standard errors in a way that takes into account correlation between states and variation due to pollsters, Wright uses a parametric bootstrap method to estimate the variance-covariance matrix for the estimates $\hat{\theta}_i$. New values for the log-ratio of percent candidate support for state i and poll j are calculated as $y'_{ij} = y_{ij} - m_k - g + \eta_k$. y_{ij} is the log-ratio from the original polls, m_k is the mean log-ratio for each pollster k within state i, g is the grand mean log-ratio for state i, and $\eta_k \sim N(0, s_m)$, where s_m is the sample variance of all the m_k values within state i. For each pollster k in state i, a new set of data is generated by bootstrapping the paired values $\{y'_{ij}, t_j\}$, which are combined across pollsters to obtain a new dataset for state i. These datasets for each state are then combined so that bootstrapped estimates of $\hat{\theta}_i(0)$ can be obtained, and the process is repeated so that 500 estimates of $\hat{\theta}_i(0)$ are calculated. Taking the covariance of these estimates then results in an estimated variance-covariance matrix, from which standard errors can be drawn.

Super Tuesday 2020 Data

Raw polling data for the 2020 Democratic primary was downloaded from FiveThirtyEight's poll tracking site. The focus of this analysis is the 14 states that voted in Democratic presidential primaries on March 3, 2020, known as Super Tuesday: Alabama, Arkansas, California, Colorado, Maine, Massachusetts, Minnesota, North Carolina, Oklahoma, Tennessee, Texas, Utah, Vermont and Virigina. (The U.S. territory American Samoa also voted on Super Tuesday, but no territory-level polls were available for analysis.) Super Tuesday was chosen because of the large number of states voting all on one day, making predictions based on spline extrapolation from the smoothing mixed effects model more straightforward, and because the Super Tuesday states seemed to include a geographically, ideologically, and demographically diverse set of U.S. voters. Super Tuesday followed caucuses and primaries in Iowa, New Hampshire, South Carolina, and Nevada, where Mayor Pete Buttigieg, Senator

Bernie Sanders, and Vice President Joe Biden won or received significant pledged delegates. Senators Elizabeth Warren and Amy Klobuchar also gained delegates in these early contests. Within two days of Super Tuesday, both Buttigieg and Klobuchar dropped out and endorsed Biden. This left Biden and Sanders as the perceived primary frontrunners, representing the moderate and progressive wings of the Democratic party, respectively.

To compare support for Biden vs. Sanders across Super Tuesday states and predict which candidate would win each state, the log-ratio of percent support for Biden to Sanders is used as the response in the smoothing mixed effects models. State-level polls in the Super Tuesday states with end dates before March 3, 2020, were used as data. Following the process described in the previous section for selecting values for the bandwidth parameters, $\lambda_\mu = 5000$ and $\lambda_v = 100$ were found to minimize the prediction errors when comparing the estimated $\hat{\theta}_i$ to the test set. By way of comparison, the λ combinations that produced the 10 lowest RMSE values when performing the training and test set comparison (Test RMSE) are provided below, along with the RMSE values when the resulting election day estimates obtained from fitting the model with the particular λ values are compared to the true results (Data RMSE). The Data RMSE values would not be available in a real-time application of this model and are included to show that λ combinations that produce low Test RMSE values do not necessarily minimize RMSE in the final data application.

Table 1: Comparison of RMSE values for different λ combintations

λ_μ	λ_v	Test RMSE	Data RMSE
500	50	0.3125	0.3726
500	100	0.2981	0.3587
500	500	0.3019	0.3138
500	1000	0.3091	0.2958
1000	50	0.3033	0.3948
1000	100	0.2910	0.3717
1000	500	0.3012	0.3058
1000	1000	0.3103	0.2889
5000	50	0.2933	0.4569
5000	100	0.2868	0.4210

Results

Figure 1 shows $\hat{\theta}_i(0)$, the estimated log-ratio of percent support for Biden to percent support for Sanders on Super Tuesday in each voting state, along with 95% confidence intervals obtained using standard errors from the bootstrapped variance-covariance matrix and the true log-ratio from election results. Estimates greater than 0 indicate that the model predicts a Biden win while estimates less than 0 indicate that the model predicts a Sanders win. The model thus predicts Biden winning Alabama, Arkansas, Massachusetts, North Carolina, Oklahoma, Tennessee, Utah, and Virginia and Sanders winning California, Colorado, Maine, Minnesota, Texas, and Vermont. The model miscalled Maine, Minnesota, and Texas, which were won by Biden, and Utah, which was won by Sanders. For 13 out of the 14 states, all but Utah, the log-ratio results are contained within the 95% confidence intervals from the model, giving the model a coverage rate of 92.86% in this setting, which is fairly close to the expected 95%. The confidence intervals for 9 of the 14 states – California, Colorado, Maine, Massachusetts, Minnesota, Oklahoma, Tennessee, Texas, and Utah – include 0, indicating that a win by either candidate was plausible, according to the model.

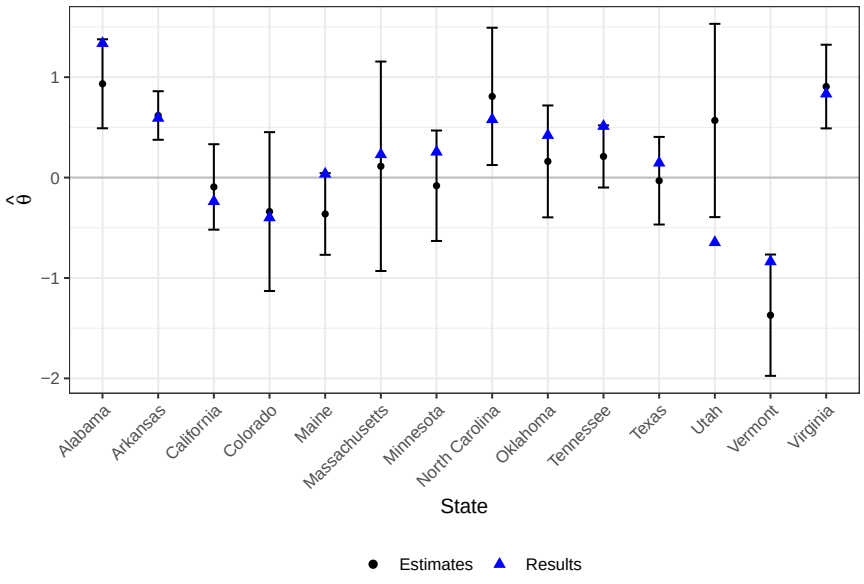

Figure 1: Log-ratio estimates for Super Tuesday states

The RMSE between the log-ratio estimates and results is 0.4210. The states where the estimates most closely match the results are Arkansas (difference in log-ratios of 0.0260), Colorado (0.0600), and Virginia (0.0722) whereas Utah has the largest difference between estimates and results at 1.2149. In terms of variation in estimates, the median standard error is 0.2534, with Arkansas having the lowest standard error at 0.1234 and Massachusetts having the highest at 0.5322.

The log-ratio results can be exponentiated to examine results on the more intuitive ratio scale of percent support for Biden over percent support for Sanders, as shown in Figure 2. On the ratio scale, the RMSE is 0.5375. Wright notes that the log-ratio results can be transformed back to the spread (or difference) scale using the formula

$$\hat{\theta}_{\text{spread}} = \frac{e^{\hat{\theta}-1}}{e^{\hat{\theta}+1}} \times 100\%$$

This relationship depends on the sum of support between the two candidates being approximately equal to 100%, however, which was not the case for these primary results. The sum of support for Biden and Sanders in the Super Tuesday states averaged around 66 percent and ranged from the low 50s to high 70s. Therefore, the estimates were not transformed to support for Biden minus support for Sanders.

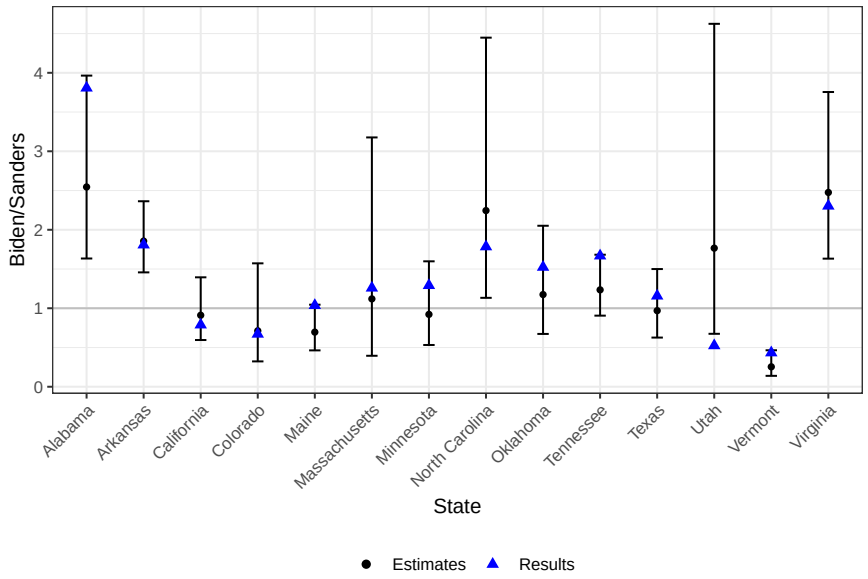

Figure 2: Ratio estimates for Super Tuesday states

Examining plots of the prediction curves for each miscalled state overlaid on the log-ratios of state polls, model estimates, and results – as shown in Figure 3 – provides some indication of why the model predicted a win for the opposite candidate. In Maine, late polls appear to fail to capture the support Biden garnered on election day, instead showing that although support for Sanders may have been waning, the state appeared to be a safe win for Sanders. In Minnesota, late polls showed decreasing support for Sanders, which was then reflected in the model's prediction curve. Sanders remained ahead of Biden in late polling, however, influencing the model estimate of a Sanders win. Late polling also may have failed to capture the effect of candidates dropping out shortly before Super Tuesday, which would lead to inaccurate estimates if the voters who supported those candidates then switched their support late to a remaining candidate. This could have been a particular issue in Minnesota, the home state of Klobuchar, who dropped out and endorsed Biden the day before Super Tuesday. Texas had substantially more state polls than the other three miscalled states, and its late polls give conflicting indications of which candidate had more support. More of these polls appear to give the edge to Sanders, however, contributing to the very slim Sanders win predicted by the model. Lastly, Utah's relative lack of state polling may have contributed to

the model's missed call. While the first two Utah polls showed strong support for Sanders, the final two indicated a steep drop in that support, with the last poll showing a Biden lead. The model predicted this swing toward Biden would continue, vastly underestimating the support for Sanders on election day.

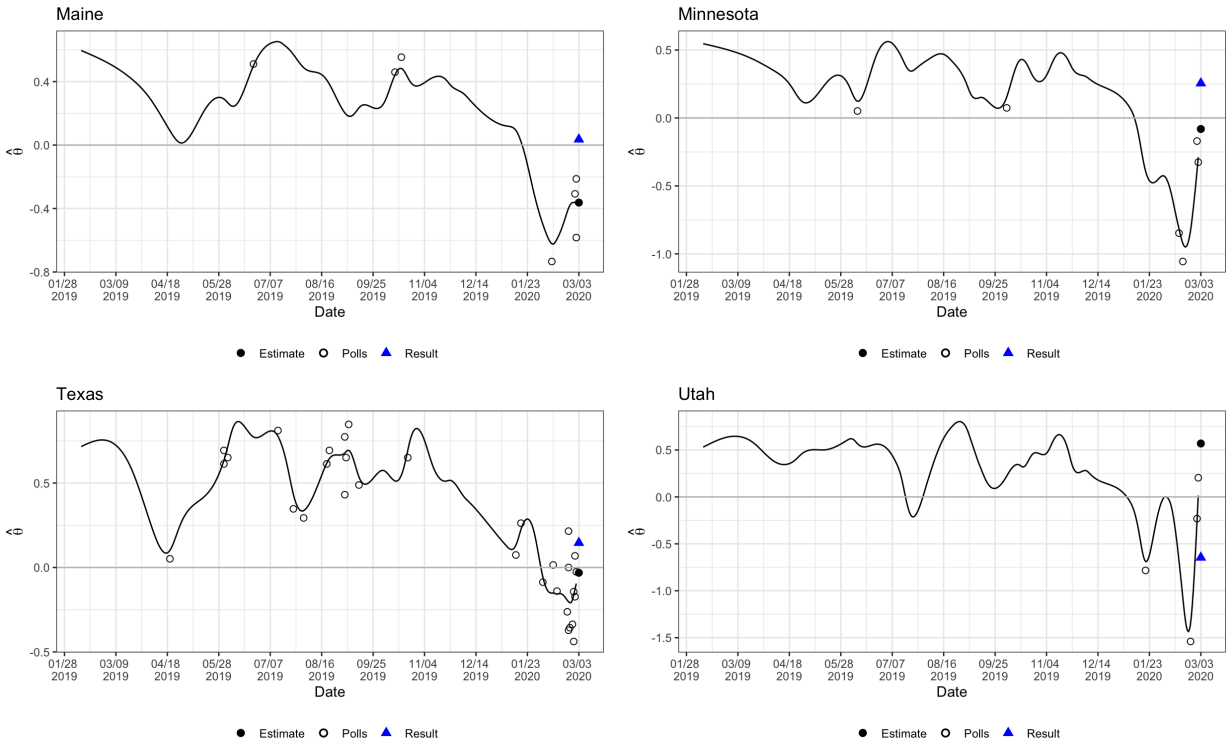

Figure 3: Prediction curves for miscalled states

One of Wright's diagnostics for the smoothing mixed effects model was to regress the model predictions on the true results; i.e., $\hat{\theta}_i(0) = \beta_0 + \beta_1 \theta_i(0) + \epsilon_i$. β_0 could then be interpreted as the model's underlying bias toward one candidate over another, and β_1 as the degree of "prediction shrinkage" for the model, with a β_1 of less than 1 indicating that the model predicts a less extreme log-ratio between the candidates, on average, than the actual results. For the Super Tuesday application, $\hat{\beta}_0 = -0.01169$ and is not significant, indicating virtually no bias in the model, and $\hat{\beta}_1 = 0.78162$, indicating that the log-ratio estimates are less extreme than the results. The correlation between log-ratio estimates and results is 0.7441. Figure 4 shows the log-ratio for each state in relation to the solid regression line (with the

dotted line representing the ideal result of $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$). States in the first and third quadrants were correctly called for Biden and Sanders, respectively, while states in the second and fourth quadrants were incorrecly called for Sanders and Biden, respectively.

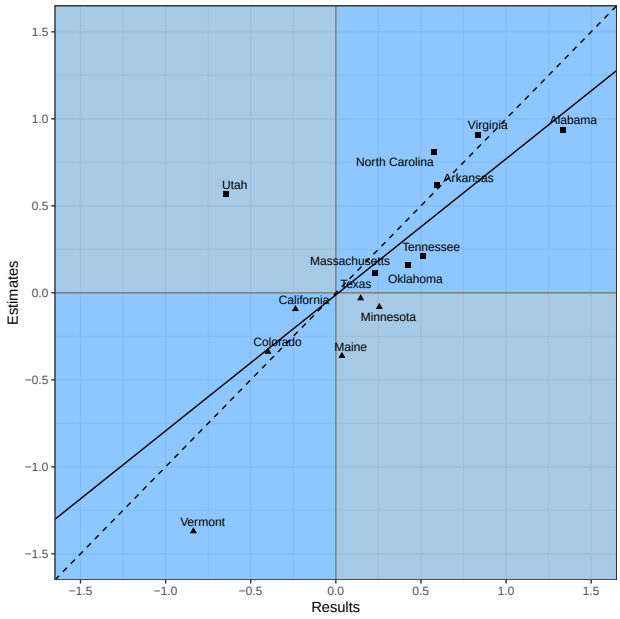

Figure 4: Regression of estimated log-ratios on true log-ratios

Because Utah appears to be a possible high leverage observation, it is worthwhile to see how the regression changes if Utah is excluded. The coefficients become $\hat{\beta}_0 = -0.16593$ and $\hat{\beta}_1 = 1.04669$. The intercept is also now significant at the $\alpha = 0.1$ level, indicating slight model bias toward Sanders. The slope being slightly above 1 indicates that the estimated log-ratios are slightly more extreme than the results. As a note, the $\hat{\beta}_1$ Wright estimated after applying the smoothing mixed effects model to the 2016 general election was 0.95, and the $\hat{\beta}_1$ for other prediction models Wright evaluated were generally substantially less than 1 (ranging from 0.63 to 0.86). A $\hat{\beta}_1$ greater than 1 seems to be uncommon, likely because it is more desirable for prediction models to correctly predict which candidate will win than to correctly predict the spread between candidates, leading to more conservative spread estimates from most models. Additionally, excluding Utah brings the correlation between log-ratio estimates and results up to 0.9250.

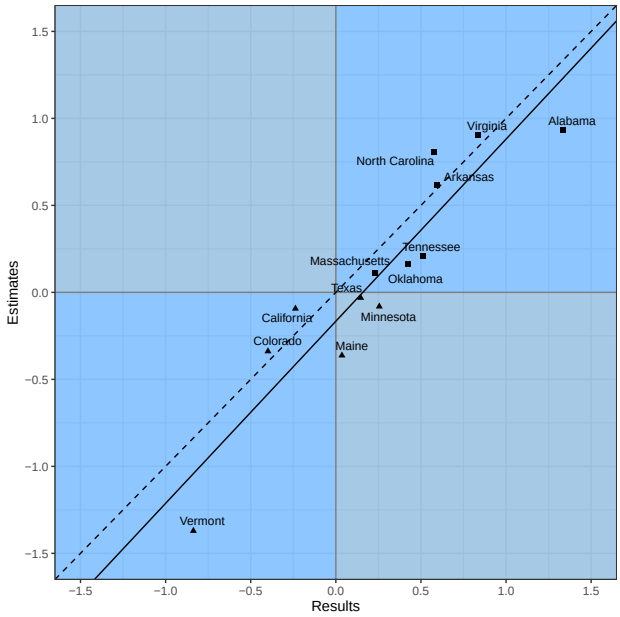

Figure 5: Regression of estimated log-ratios on true log-ratios, excluding Utah

Part of the motivation for estimating the variance-covariance matrix for the model estimates using a bootstrap method is the ability to account for correlation between states, rather than assuming all states are independent of each other. The highest correlation in magnitude between log-ratio estimates is between Vermont and Arkansas at -0.3624, followed by Vermont and Maine with 0.3408 and Virginia and Arkansas with 0.2966. This indicates fairly low correlation between the estimated log-ratios across states overall.

Because the smoothing mixed effects model uses only state polls but estimates a national trend as the fixed effect $\mu(t)$, a comparison can be made between the estimated national trend and national polls taken during the same pre-Super Tuesday period. As shown in Figure 6, the estimated national trend and its confidence bands follow the national polls well, indicating that by using polls from just 14 states, the model is able to identify a broader national trend in support for Biden vs. Sanders. (As a note, the estimated national trend begins with the mid-date of the first state poll for a Super Tuesday state, which was February 10, 2019, because the model uses only state polls.) National polls are color-coded by their FiveThirtyEight pollster rating, but there does not appear to be much of a clear

pattern between pollster rating and proximity to the estimated national trend (aside from a possible tendency of C-rated pollsters to overestimate support for Biden in the summer and fall of 2019). The RMSE between the estimated national trend and the national polls is fairly low at 0.2922.

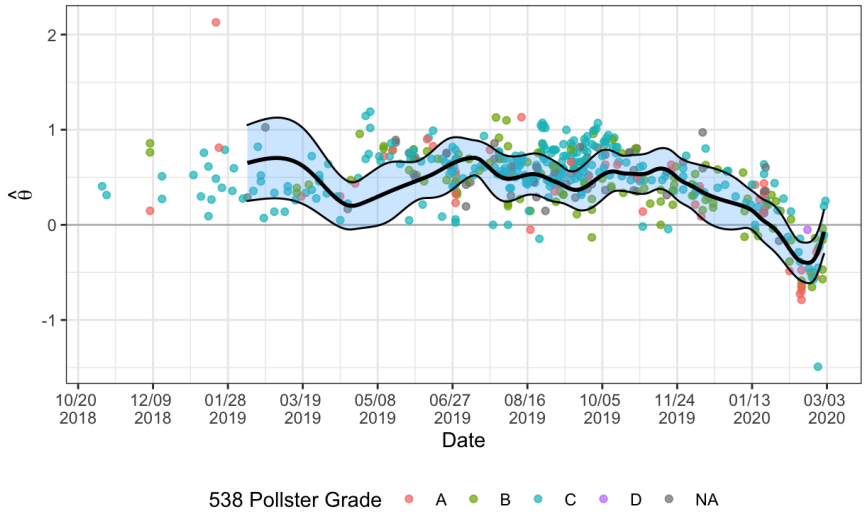

Figure 6: Estimated national trend overlaid on national polls

Lastly, the performance of the smoothing mixed effects model can be compared to the performance of FiveThirtyEight's Super Tuesday predictions. The FiveThirtyEight predictions rely on a complex model that incorporates information beyond polling data, including demographic, fundraising, endorsement, and candidate experience variables. Following metrics discussed above, FiveThirtyEight's predictions have a lower RMSE at 0.3349 and a higher correlation to the true results at 0.8688. Three states – Maine, Massachusetts, and Minnesota – were miscalled by FiveThirtyEight, with Maine and Minnesota also miscalled by the smoothing mixed effects model. As for the regression coefficients for the FiveThirtyEight predictions, $\hat{\beta}_0 = -0.13798$, indicating slightly more bias for Sanders than in the smoothing mixed effects model, and $\hat{\beta}_1 = 0.78845$, very similar to the slope found above. Overall, while FiveThirtyEight's Super Tuesday predictions performed better than those from the smoothing mixed effects model, the evaluation metrics for the two models are fairly close, showing that the models perform somewhat similarly.

Conclusion

It is notable that the smoothing mixed effects model performs well across several evaluation metrics in predicting the outcomes of 2020 Democratic presidential primaries in Super Tuesday states, given that Wright originally demostrated the strong performance of this model in the 2016 general election setting. This implies that the smoothing mixed effects model has potential as a forecasting model not only for essentially two-party elections, but also for elections with several viable candidates. A possible limitation to this finding is that the Democratic primary had practically become a two-candidate race by Super Tuesday 2020 and that the model may struggle in a setting where public support was more evenly split among three or more candidates. The Super Tuesday state where a candidate other than Biden or Sanders received the most support was Massachusetts, where Warren finished with 21.4% of the vote compared to Sanders's 26.7% and Biden's 33.6%. While the Massachusetts log-ratio estimate has the largest variation among the states, the estimate itself is close to the result, with the estimated log-ratio falling 0.1171 lower than the actual log-ratio, or an underestimate of 0.1391 on the more intuitive ratio scale. Although this is only one example, it suggests that the smoothing mixed effects model may still closely predict the outcome for the top two candidates in a multi-candidate race.

Another limitation of this application, particularly compared to Wright's general election setting, is that it is not possible to accurately convert the log-ratio estimates back to the scale of spread between candidates because the top two candidates in this primary setting did not garner near 100% total support, as is usually the case with U.S. general presidential elections. Consequently, the results of the model must be interpreted and evaluated on the log-ratio or ratio scales, which are less straightforward to understand. The model still indicates the relative support for each frontrunner in each state, however, and can be used to predict which candidate will win each state. Overall, one of the main strengths of the smoothing mixed effects model is its simplicity, relying only on state polls available before election day with

no background or historical data on each state needed. Because the model in the form that Wright outlines can only compare support between two candidates, future work would include an exploration of whether this smoothing mixed effects model could be extended or adapted to analyze relative support among three or more candidates, which would be applicable to both U.S. presidential primary and international multi-party democratic systems.

References

"FiveThirtyEight 2016 Election Forecast: Who will win the presidency?" https://projects.fivethirtyeight.com/2016-election-forecast/. Accessed 13 May 2020.

"FiveThirtyEight 2020: Who will win the 2020 democratic primary?" https://projects.fivethirtyeight.com/2020-primary-forecast/. Accessed 14 May 2020.

Hibbs, D.A., 2000. "Bread and peace voting in US presidential elections." Public Choice 104, 149–180.

Kennedy, C., et al., 2018. "An Evaluation of 2016 Election Polls in the U.S. American Association for Public Opinion Research." https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx. Accessed 13 May 2020.

Linzer, D.A., 2013. "Dynamic Bayesian forecasting of presidential elections in the states." *Journal of the American Statistical Association* 108, 124-134.

Michaelis, P., 2018. "Autoregressive effects in poll-based election models for the German federal election." *Electoral Studies* 54, 66-80.

Walther, D., 2015. "Picking the winner(s): forecasting elections in multiparty systems." *Electoral Studies* 40, 1–13.

Wright, F.A., Wright, A.A., 2018. "How surprising was Trump's victory? Evaluations of the 2016 U.S. presidential election and a new poll aggregation model." *Electoral Studies.* 54, 81-89.