# Supplementary Information

## Polls, Context, and Time: A Dynamic Hierarchical Bayesian Forecasting Model for US Senate Elections

Yehu Chen
Washington University in St. Louis

Roman Garnett
Washington University in St. Louis

Jacob Montgomery
Washington University in St. Louis

October 22, 2021

# A Summary of notations

Table A.1: Table of Notation

| | |
|---|---|
| $\mathcal{C}$ | Collection of all candidates in all races to reason about |
| $\mathcal{T}_i$ | Collection of times when opinion polls were conducted for candidate $i$ |
| $S_i$ | Number of conducted opinion polls for candidate $i$ |
| $t_{is} \in \mathcal{T}_i,\ (1 \leq s \leq S_i)$ | Time of the poll $s$ conducted for candidate $i$ |
| $n_{is},\ (1 \leq s \leq S_i)$ | Sample size of poll $s$ for candidate $i$ |
| $x_{is},\ (1 \leq s \leq S_i)$ | Number of people in poll $s$ expressing support for candidate $i$ |
| $p_{is} = x_{is}/n_{is},\ (1 \leq s \leq S_i)$ | Observed support rate in poll $s$ for candidate $i$ |
| $\mathcal{D}_i = \{t_{is}, n_{is}, x_{is}\},\ (1 \leq s \leq S_i)$ | Collection of conducted opinion polls for candidate $i$ |
| $f_i : (-\infty, 0], \to [0, 1],\ (i \in \mathcal{C})$ | Latent level of public support for candidate $i$ as a function of time, which is the sum of a linear and a non-linear components |
| $a_i + b_i t$ | Linear component in the trend of voter preference $f_i(t)$ |
| $a_i \sim \mathcal{N}(\bar{a}_i, \sigma_a^2 = 0.1^2)$ | A Gaussian prior on the intercept of linear component in $f_i(t)$ |
| $\bar{a}_i$ | Mean of the Gaussian prior on $a_i$, which is obtained from either prior knowledge or another model |
| $b_i \sim \mathcal{N}(0, \sigma_b^2 = 0.002^2)$ | A Gaussian prior on the slope of linear component in $f_i(t)$ |
| $\eta_i(t) \sim \mathcal{GP}(0, K)$ | Smooth non-linear component in the trend of voter preference $f_i(t)$ on which a zero-mean Gaussian process prior is placed |
| $K$ | Matérn covariance function with $\nu = \frac{3}{2}$ of the Gaussian process prior on $\eta_i(t)$ |
| $\rho$ | Length scale of the Matérn covariance function |
| $\lambda$ | Output scale of the Matérn covariance function |
| $V(t, t')$ | Induced covariance function after absorbing the hyperparameters controlling the priors of the linear component |
| $\sigma^2$ | General noise stemming from the polling data |
| $B_{ss} = p_s(1 - p_s)/n_s + \sigma^2$ | $S \times S$ diagonal matrix that approximates the variance in binomial likelihood plus the general noise $\sigma^2$ (dropping index for candidate $i$) |
| $\boldsymbol{\omega} = (\rho, \lambda, \sigma^2)$ | Set of hyperparameters shared across races |
| $\tau$ | Forecasting horizon |
| $m_j$ | Number of candidates in race $j$ |
| $\alpha_{ij},\ 1 \leq i \leq m_j$ | Concentration parameter for candidate $i$ in race $j$, which is a linear function of $f_i(0)$ and fundamental covariates, plus an interactive party-year random effect |
| $\gamma_{\text{year}} \sim \mathcal{N}(0, \sigma^2_{\text{year}})$ | Gaussian prior for the year random effect |
| $\tilde{\alpha}$ | Base parameter added to the concentration parameter |
| $z_{ij},\ 1 \leq i \leq m_j$ | Contextual fundamentals for candidate $i$ in race $j$ |
| $y_{ij},\ 1 \leq i \leq m_j$ | Actual vote share for candidate $i$ in race $j$ |

# B Correlation across states

In the realm of presidential forecasting, a significant amount of attention has gone into understanding how outcomes across states are correlated. The basic idea is that the vote share in, for instance, Wisconsin will often be correlated with vote share in Iowa. Thus, if we build a model assuming that outcomes across states are conditionally independent we are likely to underestimate our uncertainty.

In our model we account for this inter-state correlation using year-level random effects (interacted with party). Thus, we allow that there may be general "swings" towards one party or another in a given year, but do not otherwise consider potential correlated errors. We feel this is appropriate for the Senate elections given the structure of the data because the inter-state correlations are much lower than in the presidential race.

To further explore this issue, Figure B.1 shows the correlation between the vote share for Republican presidential candidates in each state for the 1976–2020 period. In this figure, each square represents a state dyad with darker colors indicating higher levels of correlation. We can see that outcomes are highly correlated across states and there are clearly regional and subregional patterns in the data. This indicates that even were we to control for state-level factors, there will still likely be unmodeled correlations in the residuals.

Figure B.2 shows this same analysis for Senate races in the 1976–2020 period. The figure shows that there is dramatically less correlation in outcomes across states. Moreover, once we have controlled for common factors (e.g., PVI) correlations in the residuls are even lower. Thus, while it may be possible to improve model performance by accounting for these inter-state correlations, we believe that it is not necessary in the case of Senate races.

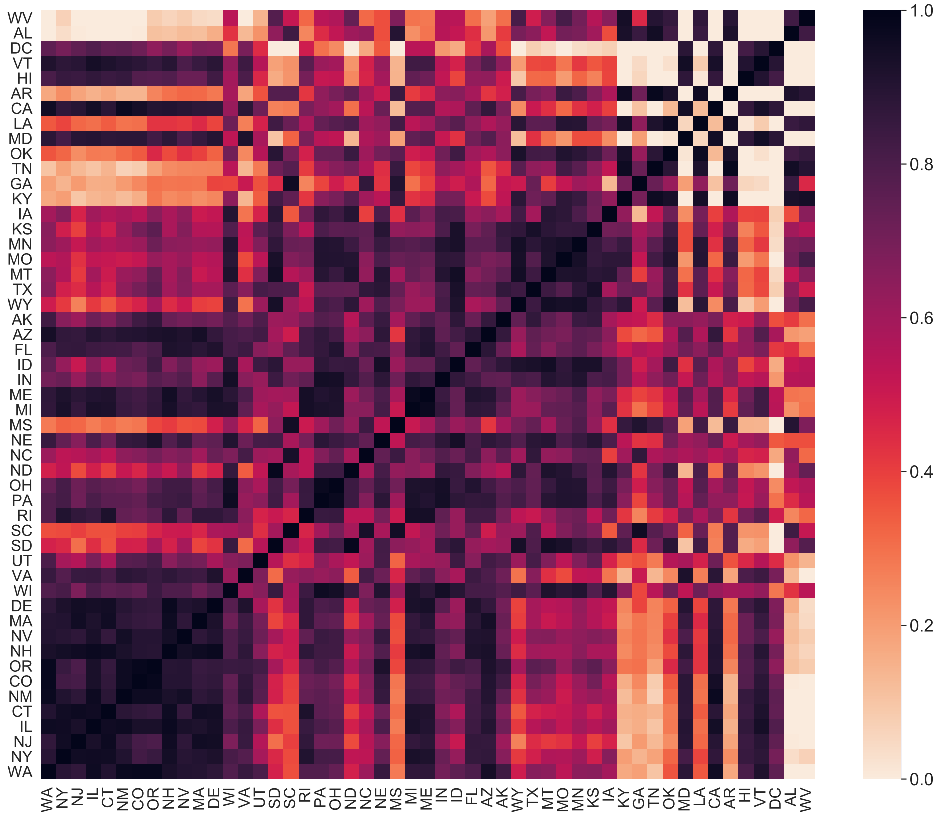


Figure B.1: Correlation between states in presidential races for Republican candidates. Correlations in outcomes are computed using election outcomes in presidential races from 1976-2020.

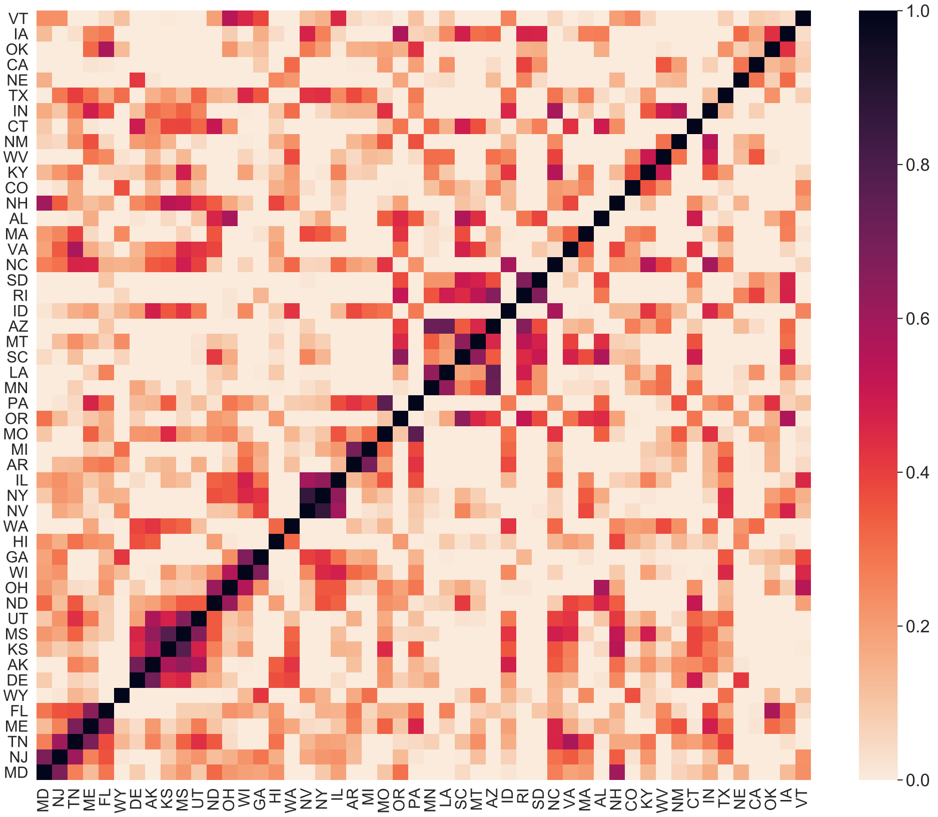


Figure B.2: Correlation between state in senatorial races for Republican candidates. Correlations in outcomes are computed using election outcomes in senate races from 1976–2020.

# C Alternative Prior Heuristic

In the main text we use a common standard deviation for the prior of 0.1. Here we briefly consider two other choices for this parameter, using the 2016 election as our test set.[1] First, we use a model where we customize the prior to be wider or narrower as the prediction moves away from the 50% threshold. Specifically, for candidate $i$ we set the prior to be one half of the absolute difference between the prior mean of the Election Day intercept $\bar{a}_i$ and 50%:

$$\frac{|\bar{a}_i - 0.5|}{2}. \tag{1}$$

This is labeled *heuristic 1* below. Second, we simply cut our parameter in half and use 0.05 (*heuristic 2*). We summarize our notations as below:

1. Main text: the standard deviation of the Election Day intercept prior is set to 0.1.

2. Heuristic 1: the standard deviation of the Election Day intercept prior is set as in (1).

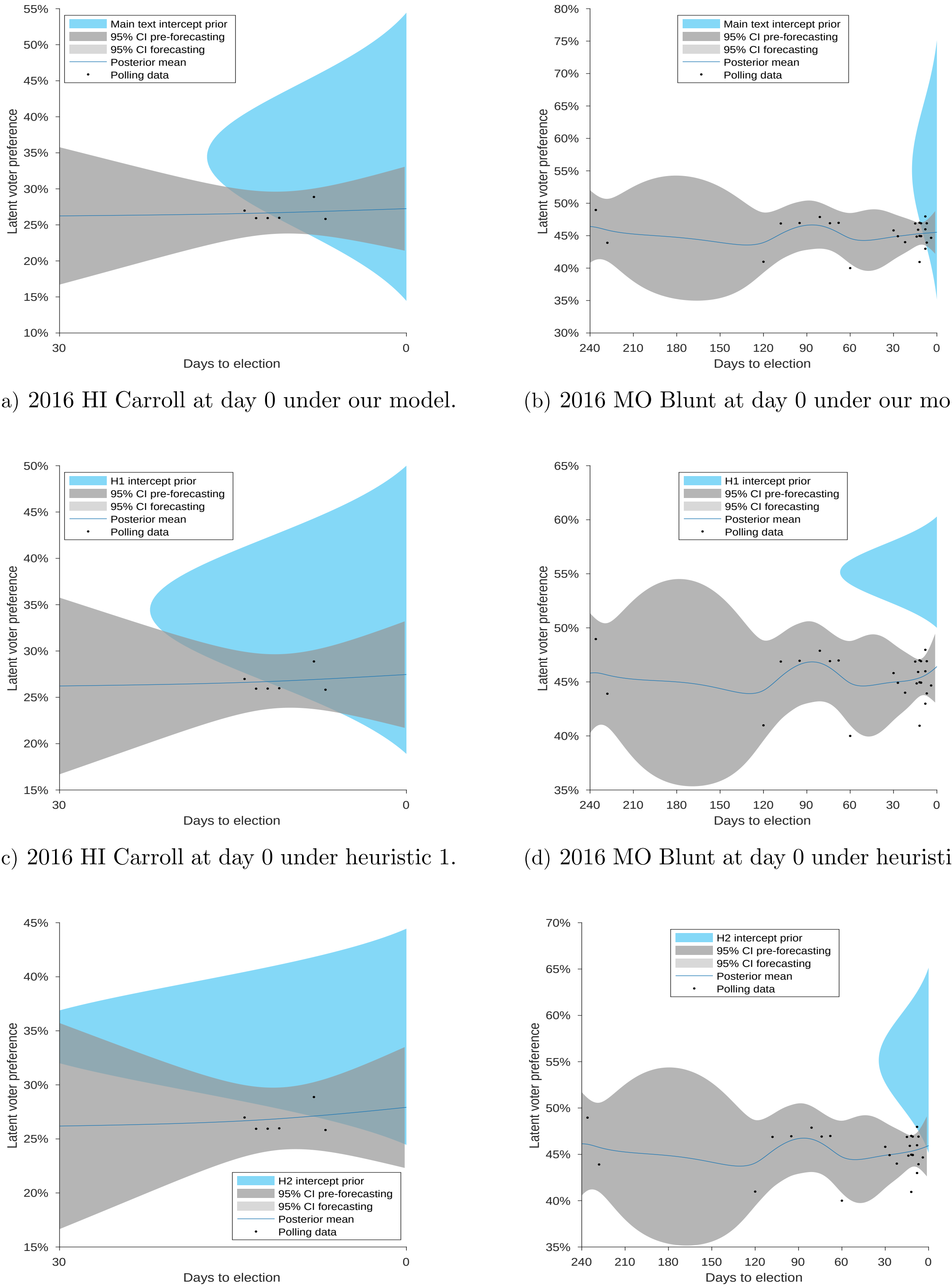
3. Heuristic 2: the standard deviation of the Election Day intercept prior is set to 0.05.

To provide a sense of how these priors work, we focus on two candidates in the 2016 election in Figure C.1. Figure C.2(a) shows the candidate level model from the main text for Caroll (HA), a lopsided race where there were few polls. Figure C.2(b) shows same for Blunt (MO), a close race where there were many polls. The remaining plots show the prior and posteriors using these alternative heuristics.

The Figures show that while the alternative priors do affect the predictions somewhat, the overall effects are very modest. To make this point clearer, Figure C.2 shows the posterior mean predictions for these races under each prior structure on the same plot. The priors do make a difference, especially for races with few polls (which are typically very lopsided). However, the differences are quite modest.

[1] Note, that at all times we try to leave the 2018 cycle as a test set and do not use it for testing out alternative variations of the model.

Figure C.1: Candidate-level models for two senate candidates under three alternative priors structures. Note that the x-axis differs for the left and right subfigures.



(a) 2016 HI Carroll at day 0 under our model.

(b) 2016 MO Blunt at day 0 under our model.

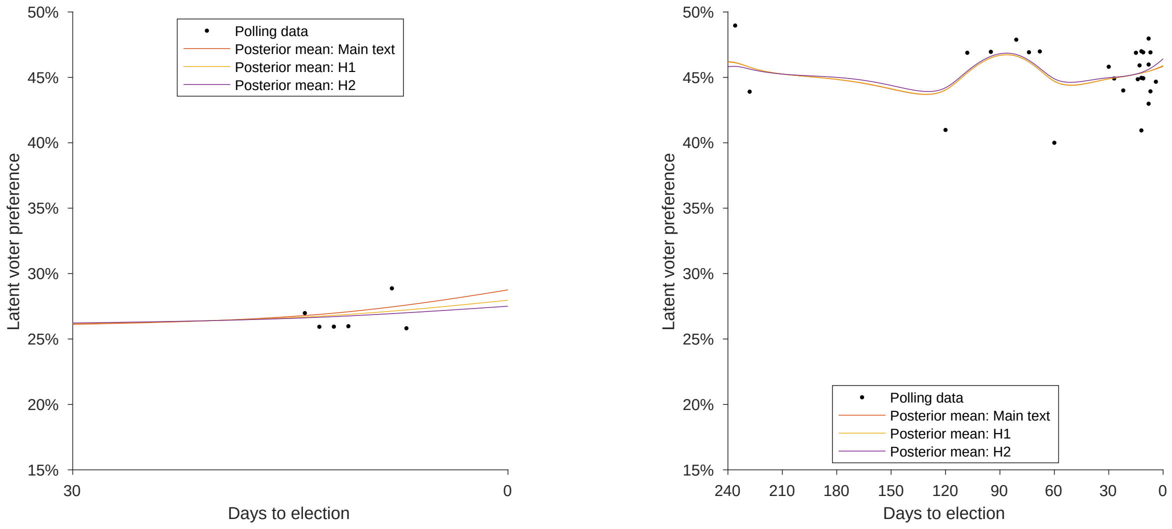
(c) 2016 HI Carroll at day 0 under heuristic 1.

(d) 2016 MO Blunt at day 0 under heuristic 1.

(e) 2016 HI Carroll at day 0 under heuristic 2.

(f) 2016 MO Blunt at day 0 under heuristic 2.

Figure C.2: Posterior mean prediction under alternative prior structures



(a) Latent voter preference posterior means of 2016 Hawaii Carroll at day 0.

(b) Latent voter preference posterior means of 2016 Missouri Blunt at day 0.

We assess how these alternatives affect the accuracy of election-level model in tables C.1, C.2, C.3, and C.4 using 2016 as our test set. Our original prior has higher averaged log likelihood, lower RMSE, and identical accuracy relative to the alternatives at almost all time horizons. At some horizons, our original prior heuristic has slightly less coverage ratio than alternatives, although this amounts to missing only one additional vote share. Our overall conclusion is that this prior does not play a decisive role in determining our forecast accuracy, but to extent it does the choice in the main text seems adequate.

Table C.1: RMSE between the posterior means and actual vote shares of 2016 races

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| Heuristic 1 | 0.0685 | 0.0682 | 0.0608 | 0.0604 | 0.0470 | 0.0397 | 0.0386 |
| Heuristic 2 | 0.0599 | 0.0598 | 0.0571 | 0.0558 | 0.0430 | 0.0366 | 0.0359 |
| Main text model | 0.0611 | 0.0593 | 0.0561 | 0.0563 | 0.0402 | 0.0363 | 0.0353 |

Table C.2: Prediction accuracies of 2016 races

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| Heuristic 1 | 0.9375 | 0.9375 | 0.9375 | 0.9688 | 0.9375 | 0.9062 | 0.9375 |
| Heuristic 2 | 0.9375 | 0.875 | 0.9375 | 0.9375 | 0.9062 | 0.9062 | 0.9062 |
| Main text model | 0.9375 | 0.875 | 0.9062 | 0.9062 | 0.9062 | 0.9062 | 0.9375 |

Table C.3: Coverage rate of actual vote shares of 2016 races

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| Heuristic 1 | 0.9688 | 0.9688 | 0.9688 | 0.9688 | 0.9688 | 0.9375 | 0.9062 |
| Heuristic 2 | 0.9688 | 0.9688 | 0.9688 | 0.9688 | 0.9688 | 0.9062 | 0.875 |
| Main text model | 0.9375 | 0.9688 | 0.9688 | 0.9688 | 0.9688 | 0.9062 | 0.875 |

Table C.4: Averaged log likelihood of actual vote shares of 2016 races (Dirichlet likelihood)

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| Heuristic 1 | 1.1738 | 1.183 | 1.3097 | 1.3136 | 1.6074 | 1.8216 | 1.8463 |
| Heuristic 2 | 1.3526 | 1.3524 | 1.5269 | 1.550 | 1.7598 | 1.8946 | 1.9166 |
| Main text model | 1.5216 | 1.599 | 1.7315 | 1.7508 | 1.8669 | 1.8974 | 1.9258 |

# D Alternative cross validation strategy

In the main text, we choose hyperparameters using a leave-one-year-out cross validation strategy. However, one concern here is that we are using future years to help choose parameters for prior years. We note that this is not a concern for our 2018 and 2020 forecasts, since we only use prior elections for these forecasts. Further, there is an advantage of using this approach since, first, we do not have to execute a new cross-validation exercise for each year in our study, and, second, we can use election cycles early in the sequence without increasing the risk of overfitting to small samples.

However, it is still worth exploring the effect of this decision further. To explore this question, we trained a model predicting the 2012 election cycle using *only* the data from 1992–2010. This included repeating the entire loyo cross-validation strategy for hyperparemeter selection. We compare these forecasts to the predictions in the main text (meaning we also include the 2014 and 2016 elections).

Table D.1 , D.2, D.3 compare the performance of these two strategies based on RMSE, accuracy rates, and coverage rates respectively. The differences are extremely minor, which suggests there is no serious issue of overfitting from the strategy we use in the main text.

Table D.1: RMSE between the posterior means and actual vote shares of 2012 races using alternative cross validation strategies.

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| Alternative CV | 0.0760 | 0.0695 | 0.0629 | 0.0609 | 0.0565 | 0.0540 | 0.0495 |
| Main text | 0.0740 | 0.0688 | 0.0624 | 0.0606 | 0.0564 | 0.0538 | 0.0494 |

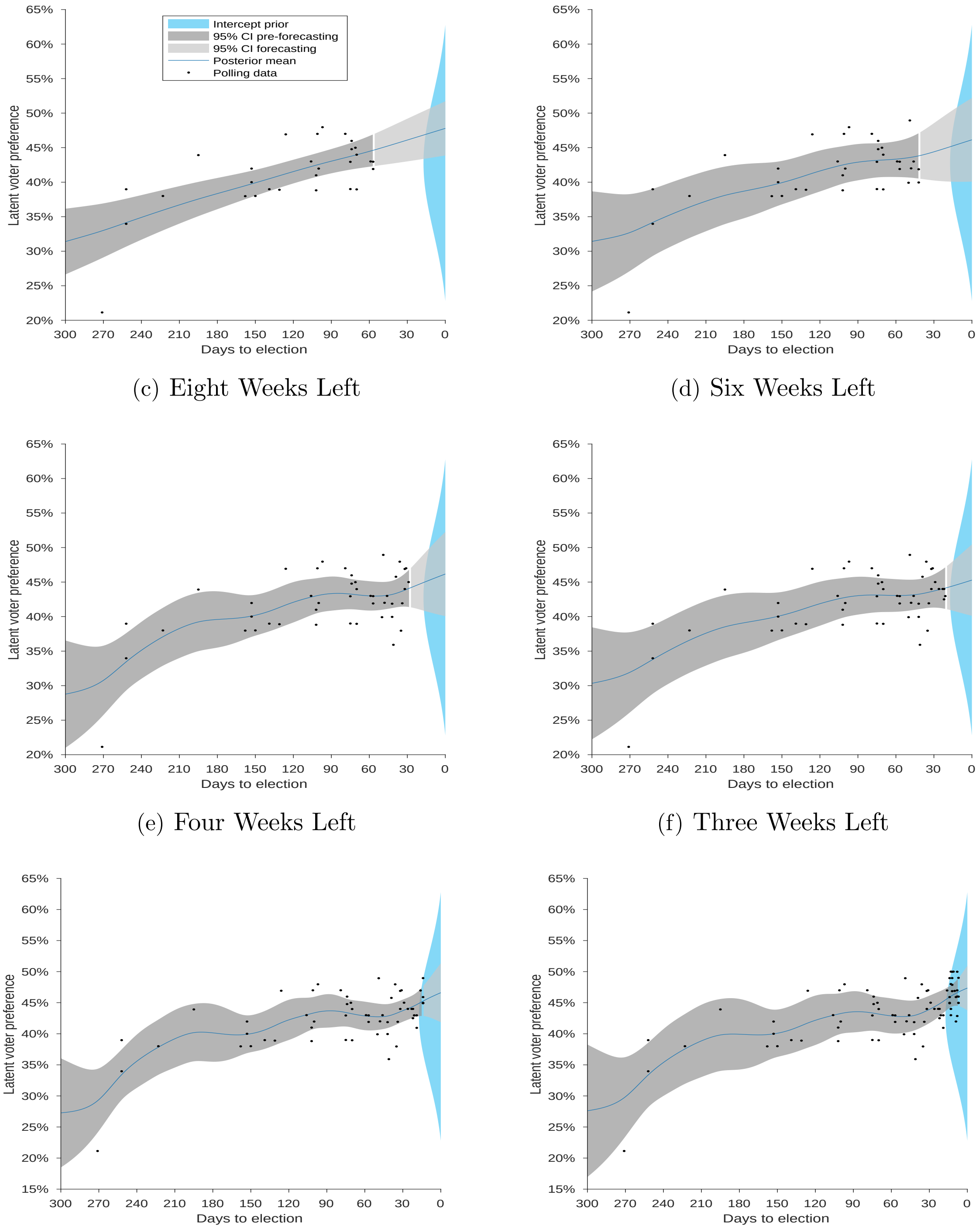
Table D.2: Prediction accuracy rates for 2012 races using alternative cross validation strategies.

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| Alternative CV | 0.8788 | 0.9091 | 0.9697 | 0.9394 | 0.9394 | 0.9394 | 0.9394 |
| Main text | 0.8788 | 0.9394 | 0.9697 | 0.9091 | 0.9091 | 0.9394 | 0.9394 |

Table D.3: 95% Coverage rates for 2012 races using alternative cross validation strategies.

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| Alternative CV | 0.9412 | 0.9412 | 0.9706 | 0.9706 | 0.9706 | 0.9853 | 0.9853 |
| Main text | 0.9412 | 0.9118 | 0.9706 | 0.9706 | 0.9412 | 0.9853 | 0.9853 |

# E Additional candidate trends: Katie McGinty (PA-2016)



(c) Eight Weeks Left

(d) Six Weeks Left

(e) Four Weeks Left

(f) Three Weeks Left

(g) Two Weeks Left

(h) One Week Left

Figure E.1: Voter Preference Estimate for Katie McGinty in 2016 Penssylvania race at various time horizons. Points represent individual polls, the blue distribution is the prior, the dark gray region is the 95% CI for the estimated latent trend, and the light-gray region is the projected latent trajectory.

# F Model parameters for Dirichlet regression fit to 1992-2016 elections

Table F.1: Summaries of Dirichlet Regression models at different horizons for 2018 model

| Horizon | Parameter | Mean | SD | 2.5% | 97.5% |
|---|---|---|---|---|---|
| $\tau = 0$ | $\tilde{\alpha}$ | 20.2 | 0.083 | 12.1 | 28.6 |
| | Voter preference | 505 | 1.11 | 408 | 609 |
| | Party:pvi | 0.78 | 0.003 | 0.52 | 1.07 |
| | Experience | 1.99 | 0.028 | 0.07 | 5.51 |
| | $\sigma_{\text{year}}$ | 5.42 | 0.028 | 3.15 | 8.50 |
| $\tau = 7$ | $\tilde{\alpha}$ | 19.0 | 0.091 | 9.62 | 28.9 |
| | Voter preference | 530 | 1.09 | 426 | 642 |
| | Party:pvi | 0.74 | 0.003 | 0.45 | 1.04 |
| | Experience | 2.14 | 0.030 | 0.07 | 6.04 |
| | $\sigma_{\text{year}}$ | 5.36 | 0.026 | 3.05 | 8.48 |
| $\tau = 14$ | $\tilde{\alpha}$ | 16.4 | 0.081 | 8.83 | 25.0 |
| | Voter preference | 442 | 1.20 | 349 | 550 |
| | Party:pvi | 0.53 | 0.003 | 0.29 | 0.79 |
| | Experience | 2.64 | 0.033 | 0.15 | 6.55 |
| | $\sigma_{\text{year}}$ | 4.05 | 0.027 | 1.90 | 6.80 |
| $\tau = 21$ | $\tilde{\alpha}$ | 10.2 | 0.084 | 2.63 | 18.5 |
| | Voter preference | 420 | 1.26 | 325 | 533 |
| | Party:pvi | 0.31 | 0.002 | 0.09 | 0.58 |
| | Experience | 2.14 | 0.031 | 0.08 | 5.91 |
| | $\sigma_{\text{year}}$ | 3.74 | 0.029 | 1.51 | 6.59 |
| $\tau = 28$ | $\tilde{\alpha}$ | 10.5 | 0.008 | 3.02 | 18.9 |
| | Voter preference | 377 | 1.24 | 281 | 485 |
| | Party:pvi | 0.26 | 0.003 | 0.05 | 0.51 |
| | Experience | 2.29 | 0.029 | 0.11 | 6.06 |
| | $\sigma_{\text{year}}$ | 2.95 | 0.037 | 0.57 | 5.64 |
| $\tau = 42$ | $\tilde{\alpha}$ | 5.50 | 0.063 | 0.44 | 12.3 |
| | Voter preference | 347 | 1.19 | 255 | 459 |
| | Party:pvi | 0.27 | 0.002 | 0.06 | 0.51 |
| | Experience | 2.79 | 0.033 | 0.16 | 6.80 |
| | $\sigma_{\text{year}}$ | 3.21 | 0.031 | 0.87 | 5.78 |
| $\tau = 56$ | $\tilde{\alpha}$ | 6.02 | 0.048 | 1.49 | 10.8 |
| | Voter preference | 205 | 0.651 | 153 | 266 |
| | Party:pvi | 0.24 | 0.002 | 0.09 | 0.41 |
| | Experience | 2.35 | 0.025 | 0.22 | 5.16 |
| | $\sigma_{\text{year}}$ | 2.73 | 0.018 | 1.27 | 4.71 |

# G Paired $t$-tests of GP+DR vs LM+DR

Here we show a paired $t$-test for the absolute errors of GP+DR and LM+DR models for predictions in the 1992–2016 period. Absolute error is defined to be the absolute distance between actual vote share and posterior median (50th percentile). Using the 1992–2016 period, here we are testing whether the absolute errors are statistically distinguishable for the two models.

Table G.1 summarizes the $p$-values and differences in mean absolute errors. At the eight and four week horizon, the test suggests no difference between GP+DR and LM+DR models. This is reasonable given that at early horizons the latent voter preferences in the GP model tend to be linear. As Election Day nears, there is a negative difference in mean (suggesting higher accuracy for the GP+DR model) and significant $p$-value. Thus, while the differences in accuracies are modest, they are statistically distinguishable.

Table G.1: Paired $t$-test for absolute errors of GP+DR vs LM+DR on the 1992–2016 races. Negative difference in mean scores are evidence in favor of the GP+DR model

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| $p$-value | $p = 0.472$ | $p < 0.001$ | $p = 0.100$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| Difference in means | 0.0096% | −0.069% | −0.043% | −0.083% | −0.104% | −0.133% | −0.149% |

# H Fit statistics for the 2018 forecast

Table H.1 shows fit statistics for predictions of the held-out 2018 cycle. The GP+DR model is accurate and well calibrated at all time horizons. The nearest competitor is the LM+DR model, which again becomes less comparable as Election Day nears and the nonlinearities become more pronounced in the candidate-level models. Note that the GP+DR model correctly predicted all but one election at $\tau = 0$ (Arizona) and all but two elections at the $\tau = 14$ horizon (Arizona and Nevada).

Table H.1: Predictive accuracy for in 2018 cycle (held out as a test set)

| | | Days until Eleciton Day ($\tau$) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Model | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
| *RMSE* | GP+DR | 0.068 | 0.063 | 0.055 | 0.054 | 0.051 | 0.047 | 0.043 |
| | GP | 0.082 | 0.081 | 0.068 | 0.067 | 0.064 | 0.060 | 0.054 |
| | LM+DR | 0.068 | 0.063 | 0.056 | 0.054 | 0.051 | 0.046 | 0.044 |
| | BRW | 0.066 | 0.065 | 0.061 | 0.060 | 0.058 | 0.053 | 0.050 |
| 95% Coverage | GP+DR | 0.958 | 0.972 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | GP | 0.917 | 0.986 | 1.000 | 1.000 | 0.986 | 0.972 | 0.944 |
| | LM+DR | 0.972 | 0.972 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | BRW | 0.681 | 0.708 | 0.653 | 0.653 | 0.681 | 0.736 | 0.681 |
| *Predictive accuracy* | GP+DR | 0.886 | 0.914 | 0.914 | 0.914 | 0.943 | 0.971 | 0.971 |
| | GP | 0.886 | 0.914 | 0.914 | 0.914 | 0.943 | 0.886 | 0.886 |
| | LM +DR | 0.886 | 0.943 | 0.943 | 0.914 | 0.914 | 0.885 | 0.943 |
| | BRW | 0.829 | 0.829 | 0.857 | 0.857 | 0.857 | 0.857 | 0.886 |
| *APLL Vote Share* | GP+DR | 1.734 | 1.861 | 2.064 | 2.063 | 2.143 | 2.285 | 2.375 |
| | GP | 1.557 | 1.670 | 1.800 | 1.814 | 1.890 | 1.990 | 2.030 |
| | LM+DR | 1.770 | 1.840 | 2.035 | 2.045 | 2.139 | 2.275 | 2.344 |
| | BRW | −1.111 | −1.394 | −2.798 | −2.629 | −3.283 | −0.319 | −0.273 |
| *APLL Winner* | GP+DR | −0.108 | −0.098 | −0.081 | −0.080 | −0.068 | −0.074 | −0.075 |
| | GP | −0.104 | −0.092 | −0.076 | −0.083 | −0.062 | −0.075 | −0.070 |
| | LM+DR | −0.108 | −0.099 | −0.082 | −0.077 | −0.071 | −0.068 | −0.068 |
| | BRW | −0.510 | −0.498 | −0.497 | −0.480 | −0.459 | −0.463 | −0.469 |

Cells reports fit statistics at various simulated time horizons using a leave-one-year-out cross validation. RMSE is root mean squared error for the point predictions, while the 95% coverage is the percent of vote shares that fall within the predicted 95% credible intervals. Predictive accuracy measures the percent of races predicted correctly across cycles. Average predicted log-likelihoods (APLL) are predicted using the Dirichlet likelihood (for vote share predictions) and the multinomial likelihood (for winner predictions).

# I Comparing 2018 forecast to 538 forecast

We can also compare our 2018 forecast to the published 538 forecast, although it is important to note that ours was not developed or released in advance of this election. Here we compare our model to the 538 forecasts for the senate at various time horizons.[2] As we note in the main text, it is not as easy to directly compare our performance to the fivethirtyeight.com forecasts as they predict non-normalized voter share (not two-party vote share), provide only 80% predictive intervals, and actually produce three predictions. Thus, for instance, the RMSE metric is not on the same scale as our model which predicts the normalized vote share (excluding write-ins, third-party votes, etc.).

Table I.1 shows the comparison in terms of RMSE, Table I.2 compares models based on prediction accuracies, and Table I.3 shows the 80% coverage rates. Almost across the board our model outperforms the various 538 models by all three metrics, although there is again evidence that our coverage rates are too conservative. Just looking at the final ($\tau = 0$) predictions, our RMSE is 0.0253, predictive accuracy is 97.14%, and coverage is 95.83%. The 538 delux model, by contrast, is 0.0374, 88.89%, and 93.94% respectively.

Table I.1: RMSE between the posterior means and actual vote shares of 2018 races

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| 538 classic | 0.0452 | 0.0409 | 0.0409 | 0.0395 | 0.0366 | 0.0347 | 0.0376 |
| 538 deluxe | 0.0455 | 0.0411 | 0.0408 | 0.0395 | 0.0367 | 0.035 | 0.0374 |
| 538 lite | 0.0411 | 0.0375 | 0.0391 | 0.0385 | 0.0353 | 0.0341 | 0.0375 |
| GP + DR | 0.0463 | 0.0406 | 0.0314 | 0.0319 | 0.0311 | 0.0253 | 0.0253 |

Table I.2: Prediction accuracy rates for 2018 races

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| 538 classic | 0.8056 | 0.8333 | 0.875 | 0.8889 | 0.9167 | 0.9167 | 0.9028 |
| 538 deluxe | 0.7917 | 0.8333 | 0.8889 | 0.875 | 0.8611 | 0.8889 | 0.8889 |
| 538 lite | 0.8889 | 0.9167 | 0.9028 | 0.8889 | 0.8889 | 0.8889 | 0.8889 |
| GP + DR | 0.8857 | 0.9143 | 0.9143 | 0.9143 | 0.9429 | 0.9714 | 0.9714 |

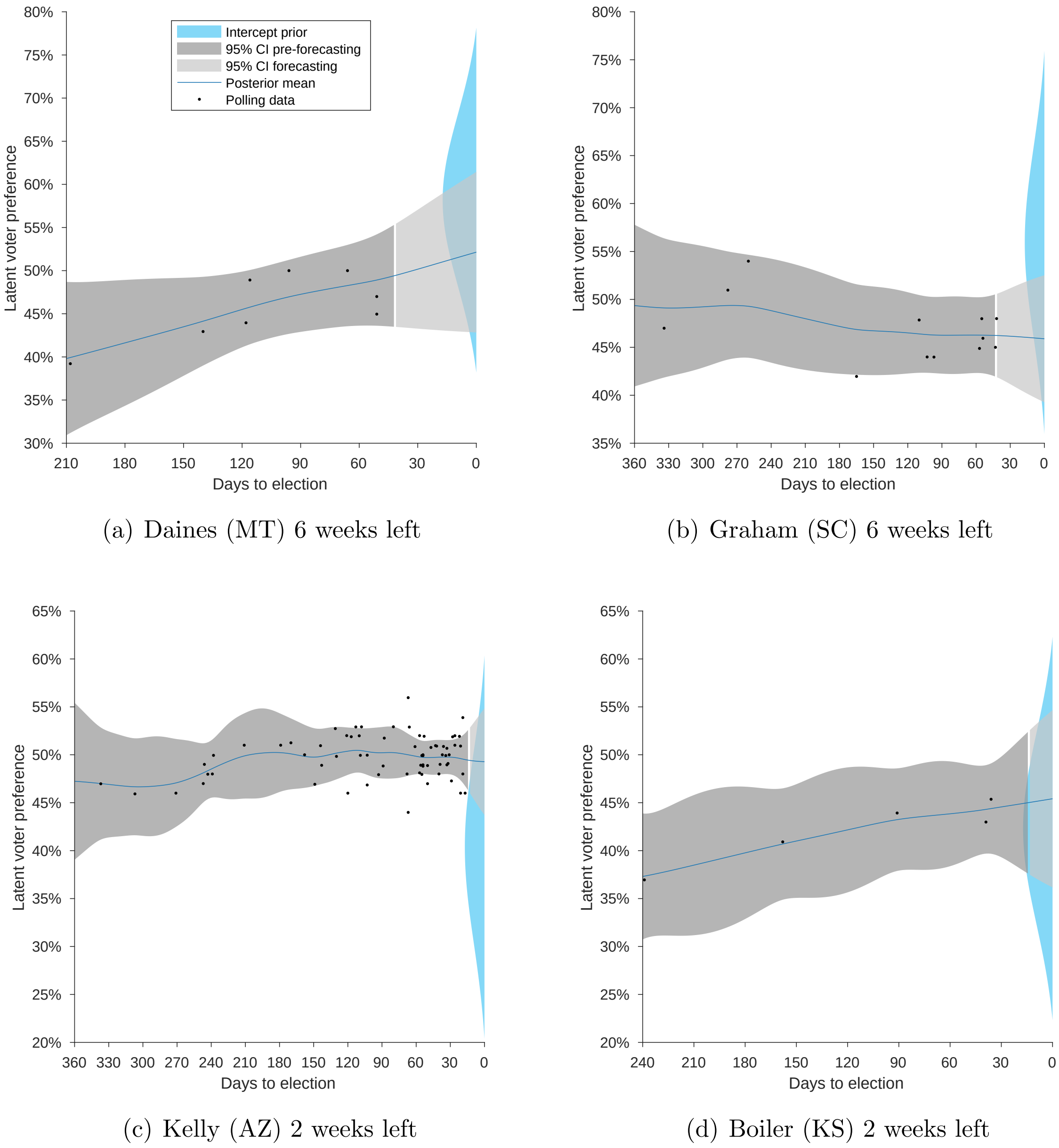
Table I.3: 80% coverage rate for forecasts at various time horizons

| Horizon | 56 | 42 | 28 | 21 | 14 | 7 | 0 |
|---|---|---|---|---|---|---|---|
| 538 classic | 0.8788 | 0.8788 | 0.8788 | 0.8788 | 0.8788 | 0.8788 | 0.9091 |
| 538 deluxe | 0.8788 | 0.8788 | 0.9091 | 0.8788 | 0.8788 | 0.8788 | 0.9394 |
| 538 lite | 0.8788 | 0.8788 | 0.8788 | 0.8788 | 0.8788 | 0.8788 | 0.9091 |
| GP + DR | 0.9306 | 0.9306 | 0.9306 | 0.9306 | 0.9028 | 1.0000 | 0.9583 |

[2] We collected these forecasts from: `https://github.com/fivethirtyeight/data/tree/master/senate-forecast-2018`

# J Example output from 2020 candidate-level models

Figure J.1 shows the results for several candidate-level models at the six-week and two-week horizons. Note that the model tends towards linearity at a farther time horizon or when the polling data is more sparse. Note also, that the prior is suggestive but does not dominate the polling data even when the horizon is still six weeks out.



(a) Daines (MT) 6 weeks left

(b) Graham (SC) 6 weeks left

(c) Kelly (AZ) 2 weeks left

(d) Boiler (KS) 2 weeks left

Figure J.1: Example 2020 candidate-level models showing trajectories of latent public opinion. Points represent individual polls, the blue distribution is the prior, the dark gray region is the 95% CI for the estimated latent trend, and the light-gray region is the projected latent trajectory.

# K Additional details for 2020 forecasts for close races

Table K.1: Median predictions and 95% credible intervals of Democratic candidates for close races in 2020

| State | 2.5 percentile | 97.5 percentile | Median |
|---|---|---|---|
| Alabama | 0.363 | 0.478 | 0.42 |
| Kentucky | 0.363 | 0.491 | 0.428 |
| Kansas | 0.394 | 0.516 | 0.455 |
| Texas | 0.404 | 0.514 | 0.459 |
| Mississippi | 0.399 | 0.522 | 0.46 |
| Alaska | 0.391 | 0.528 | 0.461 |
| Montana | 0.406 | 0.526 | 0.467 |
| South Carolina | 0.414 | 0.534 | 0.473 |
| Georgia | 0.438 | 0.543 | 0.492 |
| Iowa | 0.444 | 0.55 | 0.497 |
| North Carolina | 0.458 | 0.566 | 0.51 |
| Arizona | 0.467 | 0.573 | 0.52 |
| Maine | 0.47 | 0.579 | 0.524 |
| Colorado | 0.491 | 0.602 | 0.545 |
| Minnesota | 0.494 | 0.605 | 0.548 |
| New Mexico | 0.48 | 0.617 | 0.548 |
| Michigan | 0.494 | 0.601 | 0.549 |
| New Hampshire | 0.513 | 0.627 | 0.571 |