Bias and Excess Variance in Election Polling: A Not-So-Hidden Markov Model

Graham Tierney and Alexander Volfovsky

June 21, 2022

Abstract

With historic misses in the 2016 and 2020 US Presidential elections, interest in measuring polling errors has increased. The most common method for measuring directional errors and non-sampling excess variability during a postmortem for an election is by assessing the difference between the poll result and election result for polls conducted within a few days of the day of the election. Analyzing such polling error data is notoriously difficult with typical models being extremely sensitive to the time between the poll and the election. We leverage hidden Markov models traditionally used for election forecasting to flexibly capture time-varying preferences and treat the election result as a peak at the typically hidden Markovian process. Our results are much less sensitive to the choice of time window, avoid conflating shifting preferences with polling error, and are more interpretable despite a highly flexible model. We demonstrate these results with data on polls from the 2004 through 2020 US Presidential elections, concluding that previously reported estimates of pro-Democratic bias in 2016 and 2020 were too small, while excess variability estimates were too large.

1 Introduction

Election polls spur media discussion, inform candidate and voter choices, and provide inputs to election forecasts (Hillygus 2011). Candidates use polls to allocate campaign resources and voters may rely on polls to inform strategic decisions about who to vote for and whether to turnout to vote at all (Huang and Shaw 2009; Fey 1997; Levine and Palfrey 2007). Recent high-profile polling misses both in the US and internationally have called into question the accuracy of polls and of forecasts based on poll aggregations (Jackson 2020; Kennedy et al. 2018; Sturgis et al. 2016). Pollsters now have to grapple with declining response rates, changing methods of contact, and turbulent turnout dynamics, all of which make assessing who is being sampled and how to compare the sampled population to the expected voting population more difficult (Hillygus and Guay 2016).

Knowing that errors exist, however, does not make measuring polling errors any easier. Errors can come in two forms. Polls may suffer from a directional error, consistently over- or under-estimating one candidate's support, and excess variance, variability above what would be implied if polls were independent from each other and contacted random samples of the electorate. Estimating both of these quantities requires comparing a poll's result to the underlying value it was measuring. Making such comparisons is complicated by the fact that the "ground truth" of voters' preferences is only observed once, when the election happens, while polls measure preferences at some earlier point in time. This early measurement could be inaccurate simply due to temporal dynamics: undecided voters breaking heavily for one candidate, already-decided voters changing their opinion, or poll respondents making different turnout decisions than what they self-report to the pollster.

Standard solutions involve only using polls conducted close to the election and assuming that preferences do not change in that time window (e.g. Jennings and Wlezien 2018) or specifying a simple (linear) model for how preferences might change over time (e.g. Shirani-Mehr et al. 2018). While limiting the amount of polling data that enters the model can help these assumptions hold, it also risks results changing based on the amount of data used. We demonstrate that this does in fact happen: the conclusions of these methods are inconsistent across the subjective inclusion windows, i.e. the candidate whose support is overstated by polls changes based on how many days of polling are included. Moreover, when the assumptions about how

preferences evolve are incorrect, these methods will mislabel changes in preferences as polling errors with high precision because they do not properly account for model misspecification and the fact that the truth and measurement are observed at different times. Finally, certain implementations require modeling polling errors on the logistic scale to ensure that relevant quantities are bounded between 0 and 1. However, this complicates interpretation of estimated polling errors because they require an inverse-logistic transformation to report results with meaningful units and the directional error varies with the actual election result.

Our proposed solution to identifying biases in polls borrows from tools frequently used in the election forecasting literature (Jackman 2005; Linzer 2013). We specify a flexible hidden Markov model for how preferences change over time and treat the election outcome as a peak at the typically-hidden, underlying Markovian process. This approach directly builds upon the methodology in Shirani-Mehr et al. (2018), and has three principle advantages:

1. Consistency across time-windows: We estimate election-level errors and excess variance *without* relying on a tight time window around the election. Our model's estimates are remarkably consistent whether using data from polls conducted within 10 days of the election or within 100 days, while other models markedly shift their estimates of bias depending on the time window. The model learns and applies a weighting scheme to down-weight polls conducted far from the election, where the weights are determined by the variability of preferences: if preferences are highly variable then the information in a poll becomes outdated quickly, if preferences are stable then the information persists. This issue is highlighted in Figure 1, where we show which candidate's support is overstated by polls changes by inclusion window for traditional models but not our model.

2. Avoid conflating changes in preferences with polling error: *Polls are a snapshot, not a forecast.* If preferences are shifting, then even extremely accurate polls will appear "biased" when compared with the election result. This and the above point are especially apparent in 2008 swing states and select 2016 races with distinct non-linear trends in preferences.

3. Interpretability: Simple models for changes in errors and preferences, e.g. linear trends as in Shirani-Mehr et al. 2018, require modeling bias after a transformation of key parameters to ensure polls are measuring a quantity between 0% and 100%. Our model for preferences is flexible enough that these kind of transformations are not necessary, which enables easier interpretation, direct modeling of bias, and results that do not change based on the actual election result.[1]

We estimate that polls did not systematically over- or under-state either party across state-level contests from 2004 through 2012, while they did overstate the Democrats' support by approximately 2 percentage points in 2016 and 2020. In contrast to directly comparing the poll and election result or using a linear trend adjustment (Shirani-Mehr et al. 2018), we estimate slightly larger errors in the 2016 and 2020 elections and much smaller excess variances across all election cycles. Unlike previous approaches, our estimates are not sensitive to how many days of polling are included, letting us leverage a larger sample and estimate more credible bias and excess variability parameters. In particular, without a flexible model for preferences, other models conflate changes in preferences with polling error, estimating excess variability that is two to three times larger than what our model indicates.

2 Background

In this section we situate our methodological points within prior work on polling in Section 2.1, documenting the potential sources of non-sampling error and standard methodologies, and prior work using hidden Markov models for election forecasting and poll aggregation in Section 2.2, documenting common issues with such methods that our application avoids.

1. In particular, if error is modeled on the logistic scale, as it is in Shirani-Mehr et al. (2018), then the error of a poll on election day must be computed by initially taking the logistic transformation of the election result, adding the estimated logistic error, then taking the inverse-logistic transformation, and finally subtracting the actual election result. This number changes based on the election result because of how error is modeled. In contrast, with our method a single parameter measures polling bias directly and is invariant to changes in actual the election result. See Section 5 for details.

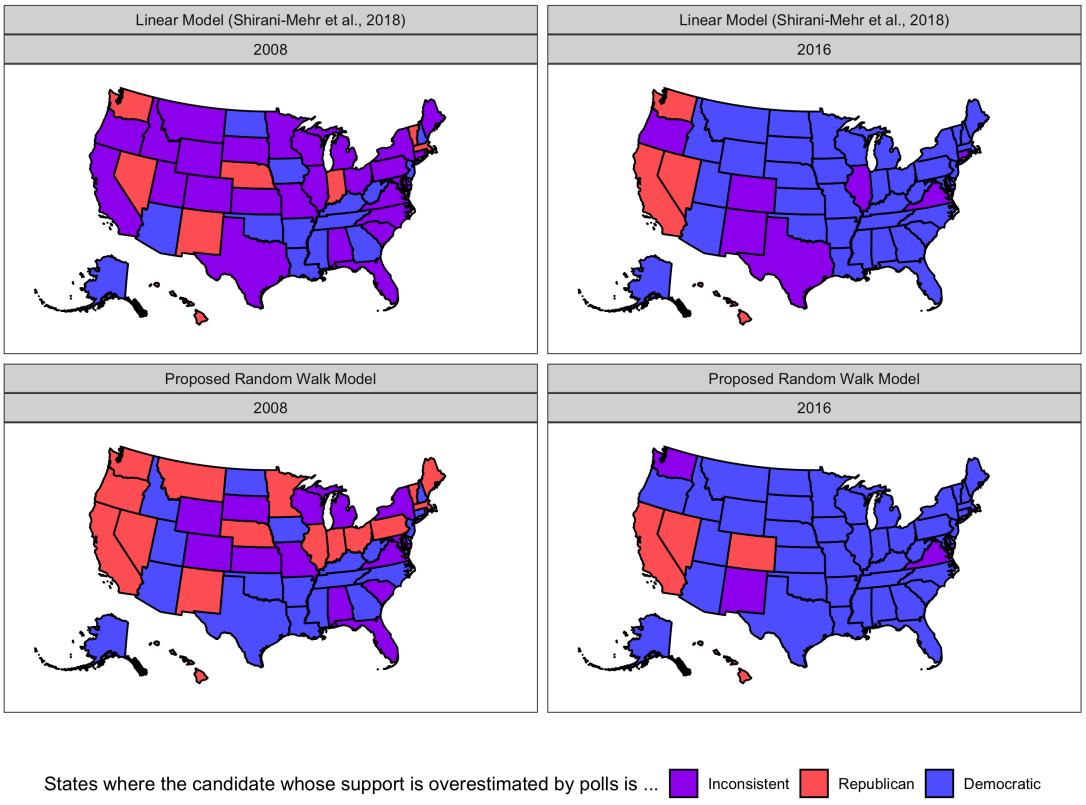

Figure 1: **Shifting Results by Inclusion Window.** Figure shows for the two most-polled election cycles in our data, 2008 and 2016, states where polling error estimated by the linear model of Shirani-Mehr et al. (2018) and our proposed model (Section 4) *changed signs*. Purple coloring indicates states where models estimate that polls are biased towards different candidates depending on which polls are included. In 2008, a very turbulent year, our model estimates much more consistent errors. In 2016, when polls overstated the Democrats support, both models are usually consistent, but our model is slightly more so.

2.1 Polling Accuracy

Pre-election or "horserace" polls have a long history in the United States. Pollsters typically attempt to contact a representative sample from the voting population (or weight a random sample to match the expected voting population) and ask respondents which Presidential candidate they intend to vote for. While errors certainly do arise from the random sampling, the extensive literature on total survey error documents many non-sampling reasons for polling errors (Weisberg 2009; Biemer 2010; Groves and Lyberg 2010). For example, non-response bias may occur when supporters of a candidate with low support may be less likely to respond to polls (Gelman et al. 2016). Other sources of error include order effects (McFarland 1981) and question wording (Smith 1987). For election polls specifically, many pollsters poll the same race and differences in survey methodology and question wording contribute to "house effects" whereby each pollster may measure preferences slightly differently. McDermott and Frankovic (2003) study house effects in the 2000 US Presidential election and Jackman (2005) study them in the Australian context. Our method studies non-sampling errors to examine how they vary across election cycles and voting populations.

After the high-profile polling misses in the state-level results of the 2016 and 2020 US Presidential elections, practitioners began to question the relevance of election polling altogether (Barnes 2016). Indeed, lengthy retrospective reports about those elections and horse-race polls were produced, which suggest that non-representative samples and potentially late swings in opinion contributed to the errors (Sturgis et al. 2016; Kennedy et al. 2017; Clinton et al. 2021).

The central theme of the literature that we build upon is that polling errors change over the course of a campaign and across different election cycles and electorates (Jennings and Wlezien 2018). As undecided voters commit to a candidate and late-breaking news stories affect voter preferences, poll results late in the campaign are generally closer to the actual election outcome. A frequent methodological choice in this literature is to compute poll-specific errors as the difference between a poll's stated support for each candidate and the election outcome for some small time window close to the election. For example, Kennedy et al. (2018) use polls within 13 days of the election and Jennings and Wlezien (2018) use those within 7 days. This structure was relaxed most recently in Shirani-Mehr et al. (2018) who allow for linear changes throughout a 21-day period before the election. We detail their method further in Section 4 as a principle comparison for the model we develop. Our model will flexibly capture shifting preferences and account for them when estimating polling error.

2.2 Forecasting and Poll Aggregation

Over the last 20 years, aggregating polls to create more precise estimates of the electorate's preferences and forecasting eventual election results has risen in popularity (Jackson 2018). We focus this section on only the specific class of models that we build upon for our method, highlighting key innovations and areas where our application simplifies certain assumptions.

Linzer (2013) developed a hidden Markov model for US Presidential election forecasting where underlying state-level preferences evolve over time following a random walk. A Bayesian estimation procedure enables forecasting Electoral College outcomes via posterior predictive simulations. Jackman (2005) outlines a very similar model that is more focused on pooling polls to estimate current preferences and house effects rather than making explicit election forecasts. Pickup and Johnston (2007, 2008) expanded upon Jackman's work to estimate house effects and industry-wide bias in the 2004 and 2006 Canadian and 2004 US Presidential elections. Our method, rather than estimating pollster-specific errors from multiple polls of the same race, will estimate aggregate errors across multiple elections.

The underlying principle, that polls are noisy measurements of latent preferences that change over time, has been expanded and applied by many forecasting models, including multi-party systems (Walther 2015; Stoetzer et al. 2019), and the Economist's 2020 forecast, which made additional adjustments for correlated shifts in state-level trends and polling errors (Heidemanns, Gelman, and Morris 2020). These models often "debias" current polls by correcting for historical polling errors (Rothschild 2009). Recent work applied this general principle to predicting US Senatorial elections with more complex mapping from polling data to election outcomes (Chen, Garnett, and Montgomery 2022). The general modeling framework whereby surveys measure latent preferences has been applied beyond election polls (e.g. Caughey and Warshaw 2015).

These methods typically incorporate "fundamentals," historical data on election outcomes, broad economic and political features, and potentially very early polls, into priors on election results (Abramowitz

2008; Campbell and Wink 1990; Erikson and Wlezien 2008). We, however, will condition on the election outcome to estimate polling errors, removing the need for forecasting and applying fundamentals approaches altogether. Another area of concern for these models is how correlated changes in preferences are across electoral units. Consistent with Shirani-Mehr et al. (2018), we allow sharing of information on the election-level parameters through a hierarchical model across elections.

Section 2.1's discussion of non-sampling sources of error highlight that polls have significant sources of unknown uncertainty, which contribute to forecasts based on those polls making overconfident and inaccurate predictions (Jackson 2020). The overconfidence is largely attributable to the fact that the (invalid) assumptions that polls are independent from each other and random samples from the electorate lead to significantly underestimating the variance of estimators that combine results from multiple polls (Clinton and Rogers 2013). While the model we use stems from the forecasting literature, we will treat the election outcome as known and use that information to derive estimates of historic polling errors and uncertainty, which can inform and improve future forecasts.

3 Data

Our data consist of the 100 day sample of polls used in Shirani-Mehr et al. (2018) covering 2004 through 2012 elections and is supplemented with 2016 and 2020 polling data for the 100 days preceding those elections collected by the Economist for use in their forecasting model.[2]

The availability of polling data varies across elections and states. Elections years and swing states where the outcome is genuinely in doubt are polled more frequently. Figure 2 shows the number of polls by election year for varying time windows before the election. The election in 2008 was the first time when an African American candidate was nominated by a major political party, and more polls were conducted that year than in any other. Figure 3 shows the number of polls conducted at most 100 days before the election by state and year. The 2008 election cycle again stands out, as do many swing states such as Ohio, Pennsylvania, and Florida.

4 Proposed and Comparison Models

The core model that we study is that poll i of election r_i conducted t_i days before the election reports y_i, the proportion of the sample that intend to vote for the Republican candidate out of people who intend to vote for either the Republican or the Democrat (omitting third-party and undecided voters), and n_i, the number of intended two-party voters. In the election, the Republican's portion of the two-party vote is v_{r_i}. Some polls are conducted years before the actual election, so all models require a cutoff day T that identifies which polls to look at, i.e. only polls with $t_i \leq T$. Below we discuss several comparison models and develop our proposed model. While all models share the assumption $y_i \sim N(p_i, \sigma_i^2)$, *they differ in how p_i, the true underlying preference in poll i, is decomposed.*

M1: Static Model. The simplest model we consider is commonly used in practice where $p_i := v_{r_i} + \alpha_{r_i}$. Writing $y_i - v_{r_i} \sim N(\alpha_{r_i}, \sigma_i^2)$, this assumes that the electorate's preferences do not change over time and α_{r_i} is a time-invariant, election-specific error. This requires choosing a small time cutoff T "close" to the election to justify the assumption of static preferences.

M2: Linear Model The model in Shirani-Mehr et al. 2018 sets $\text{logit}(p_i) = \text{logit}(v_{r_i}) + \alpha_{r_i} + \beta_{r_i} t_i$ and is our principle comparison model. The authors refer to the sum $\alpha_{r_i} + \beta_{r_i} t_i$ as "error." An equivalent interpretation that clarifies the comparison with our model is to interpret $\text{logit}(v_{r_i}) + \beta_{r_i} t_i$ as preferences that change over time, and thus α_{r_i} is the only "error" term. "Error" is used here rather than bias because, due to the logistic transformation, the α terms do not measure bias in the statistical sense. Note that $t_i = 0$ corresponds to a poll conducted on the day of the election, so regardless of whether one thinks of $\beta_{r_i} t_i$ as preferences or error, the "election day error" is measured with α_{r_i} alone. Also note that the logit transform is necessary to ensure that p_i lies between 0 and 1. A linear trend could easily imply that the expected poll proportion is outside of $[0, 1]$.

2. Data are available at the replication link in the Data Availability section.

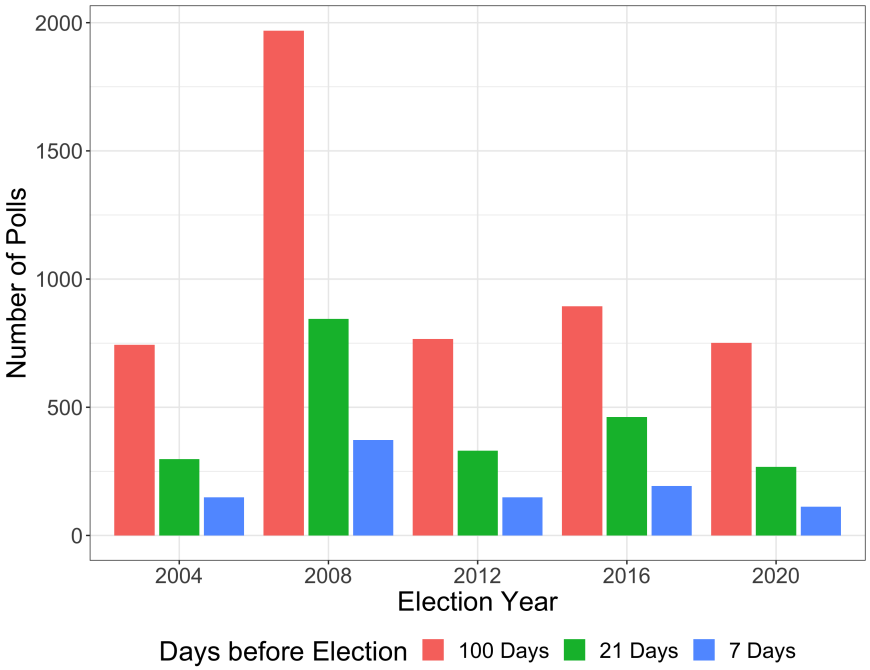

Figure 2: **Number of Polls by Varying Time Window.** Columns show the number of polls conducted each election cycle; colors indicate the cutoff time. Polls are conducted more frequently towards the end of the election cycle; approximately 19% and 43% of polls conducted in the final 100 days are conducted in the final 7 and 21 days, respectively.

M3: Random Walk (RW) Model. This is our proposed model where we set $p_i = \theta_{r_i,t_i} + \alpha_{r_i}$. θ_{rt} represents the electorate's preferences at time t and evolves via a reverse random walk: $\theta_{r,t+1} \sim N(\theta_{rt}, \gamma_r^2)$. The election result is the reveal of $\theta_{r,0} := v_r$. Note the lack of a logit transform on p_i. While one could include it, the flexibility of the random walk and small estimated γ_r terms mean that in practice neither the parameter θ_{rt} nor $\theta_{rt} + \alpha_t$ come close to leaving the interval $[0, 1]$.[3] A detailed discussion of this model is presented in the next section.

Variance terms. The difference in the above models is in the specification of each poll's expected value p_i, but the variance term σ_i^2 deserves discussion as well. The key modeling decision is whether to include the binomial variance term and, if so, how to allow for excess variance. If a poll is truly a random sample, then $\text{Var}(y_i) = p_i(1 - p_i)/n_i$. It is well known, however, that election polls have higher variance than what this would imply, despite polling firms typically constructing error estimates with this assumption. Consistent with Shirani-Mehr et al. 2018, we model this term as $\sigma_i^2 = \frac{p_i(1-p_i)}{n_i} + \tau_{r_i}$. Additive excess variance is preferable to multiplicative variance because a poll's error cannot be shrunk to essentially zero with large enough sample size. National and online-only polls can have quite large n_i, which makes the binomial variance shrink to near zero even for $p_i = 0.50$.

4.1 Random Walk Model Construction

The Static Model is quite simple and the Linear Model is detailed extensively in Shirani-Mehr et al. 2018. Here we provide additional details and expand on the interpretaiblity of our RW Model (M3). For clarity,

3. When such models are used for forecasting, the logistic transformation is more common, as in Linzer 2013, because v_r is unknown and forecasts are made with long time horizons.

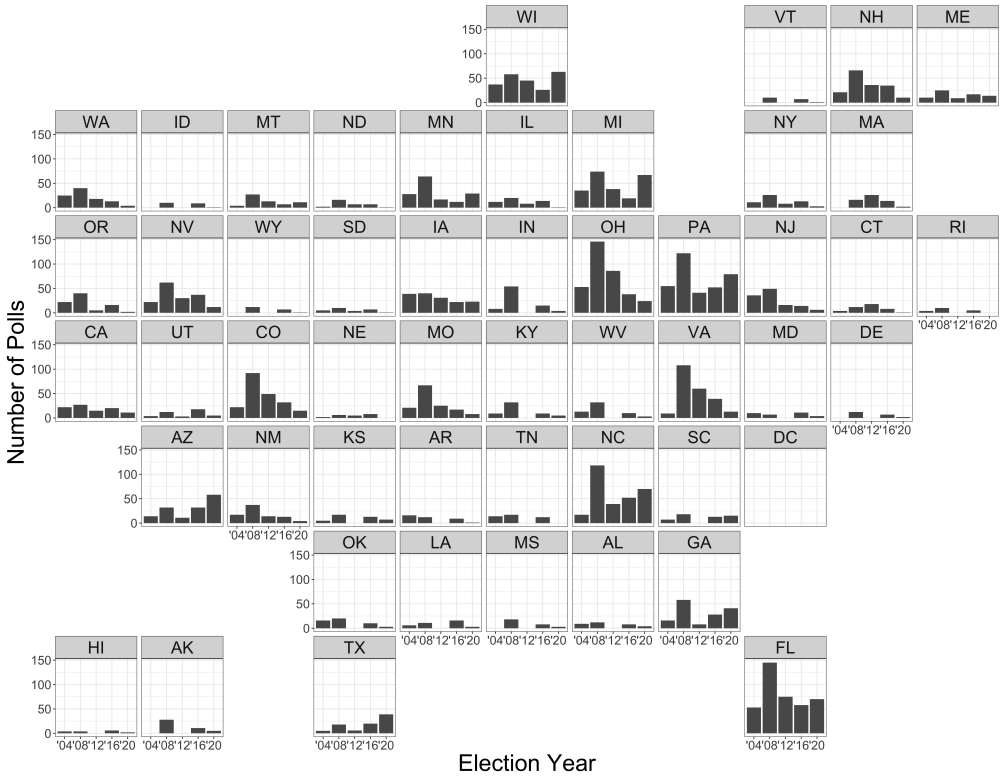

Figure 3: **Number of Polls by State and Year.** Columns show the total number of polls conducted within 100 days of the election for each election cycle in each state. Swing states are polled much more frequently than non-swing states, and many states are not polled at all in some election cycles.

we formally state the model.

$$y_i \sim N\left(p_i, \frac{p_i(1-p_i)}{n_i} + \tau_{r_i}^2\right) \tag{1}$$

$$p_i = \min(\max(0, (\theta_{r_i,t_i} + \alpha_{r_i})), 1) \tag{2}$$

$$\theta_{r,t+1} \sim N(\theta_{r,t}, \gamma_r^2) \tag{3}$$

$$\theta_{r,0} := v_r, \tag{4}$$

where τ_r is the election-specific variance above simple random sampling.

Similarly, γ_r measures how much the electorate's preferences change day to day. Under this model specification, approximately 95% of daily shifts will be $\pm 2\gamma_r$ percentage points. Lastly, α_r directly measures poll bias. A poll conducted on election day ($t_i = 0$) would in expectation be off by α_r percentage points. The minimum and maximum operators in (2) ensure that p_i lies in the interval 0 to 1 so that the variance in (1) is always non-negative.[4] On the other hand, in the Linear Model, a poll conducted on election day will have expected value $\text{logit}^{-1}(\text{logit}(v_r) + \alpha_r)$, and as such will have "election day error" of $100(\text{logit}^{-1}(\text{logit}(v_r) + \alpha_r) - v_r)$ percentage points. Note that this error changes depending on the actual election result.

Election-specific scalar parameters τ_r, α_r, and γ_r have hierarchical normal or half-normal (for variance terms) priors placed on them to borrow strength across elections. $\alpha_r \sim N(\mu_\alpha, \sigma_\alpha^2)$, $\tau_r^2 \sim N_+(0, \sigma_\tau^2)$, and $\gamma_r \sim N_+(0, \sigma_\gamma)$. We use the same "weakly informative" priors on the hyperparameters of those distributions as in Shirani-Mehr et al. 2018 for α_r and τ_r^2: $\mu_\alpha \sim N(0, 0.05^2)$[5], $\sigma_\alpha \sim N_+(0, 0.2^2)$, $\sigma_\tau \sim N(0, 0.05^2)$. For

4. In practice, this restriction is only necessary so that early (pre-convergence) MCMC draws do not break the sampler. Posterior samples are never observed close to this boundary condition.

5. In Shirani-Mehr et al. 2018, the standard deviation is set to 0.2 because α is defined on the logistic scale and the desired range is 0 ± 5 percentage points. We can directly set that prior.

γ_r, we use $\sigma_\gamma \sim N_+(0, 0.01^2)$, which places nearly all prior mass on preferences changing by at most ± 2 percentage points. The model is estimated in Stan using Hamiltonian Monte Carlo (Stan Development Team 2020).[6]

4.2 Use of observed errors $y_i - v_r$

In this section, we discus how each model estimates α_r as a function of observed polling errors $y_i - v_r$. Consider a single election r and all of its corresponding polls, $\mathbf{Y}$, conducted at least T days before the election. For clarity of exposition, assume that n_i is sufficiently large such that the binomial component of the variance $(p_i(1-p_i)/n_i)$ is negligible. A nice feature of the Static Model is that one can easily see how $y_i - v_r$ is used in estimation: under the Static Model, the posterior of α_r with a normal prior $\alpha_r \sim N(\mu_\alpha, \sigma_\alpha^2)$ is:

$$\alpha_r|\mathbf{Y}, M_1 \sim N\left(\frac{n_r/\tau_r^2}{n_r/\tau_r^2 + \sigma_\alpha^{-2}} \frac{\sum_i (y_i - v_r)}{n_r} + \frac{\sigma_\alpha^{-2}}{n_r/\tau_r^2 + \sigma_\alpha^{-2}}\mu_\alpha, \left(n_r/\tau_r^2 + \sigma_\alpha^{-2}\right)^{-1}\right),$$

where $\mathbf{Y}$ is all the polls of election r and n_r is the number of election r polls. With $\mu_\alpha = 0$ and σ_α^2 sufficiently large, this posterior simplifies to have expected value of $\sum_i (y_i - v_r)/n_r$, an equally weighted average of the observed differences between the polls and election outcome. Clear in this construction is the importance of T under the Static Model. All polls are weighted equally, regardless of when they were conducted, so T must be carefully chosen.

The Linear Model does not have this interpretation even when $\beta_r = 0$. Recall that under this model for a poll with large sample size: $y_i \sim N(p_i, \tau_r^2)$ where $p_i = \text{logit}^{-1}(\alpha_r + \beta_r t_i + \text{logit}(v_r))$. This likelihood is not log-quadratic in α_r, so it is not conjugate with a normal prior. The logistic transformation required to ensure polls' expected values lie between 0 and 1 means that the estimated α_r cannot be expressed as a function of the observed errors $y_i - v_r$.

The proposed Random Walk Model does have an intuitive use of observed polling errors. First, consider a single poll, y_i. Note that θ_{rt_i} has marginal distribution $N(v_r, t_i\gamma_r^2)$ when integrating out θ_{r1} through θ_{r,t_i-1}, so $y_i \sim N(v_r + \alpha_r, t_i\gamma_r^2 + \tau_r^2)$. Thus, the posterior for α_r with the same normal prior as above will be:

$$\alpha_r|y_i, \tau_r, \gamma_r, M_3 \sim N\left(w_i(y_i - v_r) + (1 - w_i)\mu_\alpha, \lambda_i^{-1}\right),$$

with $\lambda_i = (\tau_r^2 + t_i\gamma_r^2)^{-1} + \sigma_\alpha^{-2}$ and $w_i = (\tau_r^2 + t_i\gamma_r^2)^{-1}/\lambda_i$. Thus, w_i is the weight given to the observed error and $1 - w_i$ the weight given to the prior mean. A poll farther from the day of the election (larger t_i) will have less weight than one close to the election. As the electorate's preferences become more variable (larger γ_r) polls get down-weighted more the farther out they are conducted.

These weights highlight important improvements in our model over the Static and Linear Models. If preferences are fairly constant (γ_r is small), then polls early in the campaign still provide accurate information about α_r and correspondingly w_i is still large even for large t_i. If preferences are highly variable (γ_r is large), then early polls do not provide much information and w_i will be small for large t_i. The Static Model makes no account for the time a poll was conducted, and the Linear Model only accounts for time with a linear trend, which does not have this dynamic weighting structure. When either alternative model's key assumption about how preferences evolve hold, our model is flexible enough to recover that same structure. When those assumptions are incorrect, our model adapts to the data and still produces valid estimates.

We can derive analogous results when multiple polls are conducted. Integrating out θ_{rt} will induce dependence between the polls. The data contribution to the posterior mean is still a weighted average of $y_i - v_r$ across i, but the weights are more complex than just observation error and time until the election because of the dependency. Recall that $\mathbf{Y}$ contains all polls y_i of election r and t_i denote the number of days before the election that poll i was conducted: $\mathbf{Y} \sim N(v\mathbf{1} + \alpha\mathbf{1}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \tau^2\mathbf{I} + \gamma^2\mathbf{T}$ and $T_{ij} = \min(t_i, t_j)$. Under $p(\alpha) \propto 1$ we have

6. To improve convergence, we reparameterize the model to let $z_{rt} := \theta_{rt} + \alpha_t$, sample the posterior of z_{rt} and α_r, then use those samples to recover θ_{rt}. This is analogous to the centered parameterization described in Prado and West (2010).

$$
\begin{aligned}
p(\alpha|\mathbf{Y}) &\propto \exp\left(-\frac{1}{2}(\mathbf{Y}-v\mathbf{1}-\alpha\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-v\mathbf{1}-\alpha\mathbf{1})\right) \\
&\propto \exp\left(-\frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}{2}\left[\alpha-(\mathbf{Y}-v\mathbf{1})'\boldsymbol{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})\right]^2\right) \implies \\
\alpha|\mathbf{Y} &\sim N\left((\mathbf{Y}-v\mathbf{1})'\boldsymbol{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}), \left(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\right)^{-1}\right)
\end{aligned}
$$

That is, the posterior mean of α is a weighted average of the observed polling errors, $y_i - v_r$ where the weights are determined by the time the poll was conducted and the variability of preferences. Note that if $\gamma_r = 0$, then $\boldsymbol{\Sigma} = \tau_r^2\mathbf{I}$ and the result matches the Static Model.

5 Comparisons

We compare the models based on election day error (α_r for the Static and Random Walk (RW) Models and $\text{logit}^{-1}(\text{logit}(v_r) + \alpha_r) - v_r$ for the Linear Model) and excess margin of error ($2\tau_r$, the margin of error for poll large enough that the binomial variance is negligible). For election day error, positive values indicate the Republican candidate's support is overstated by polls. We estimate each model 10 times including only polls at most 10, 20, ... 100 days before the election with each estimation.

Figure 4 highlights two elections where the three models give different results. In the 2008 election in Pennsylvania (left panels of Figure 4), many early polls showed large Republican support and large swings are evident in the 100-day period before the election. Under our proposed RW Model, election day error and excess margin of error point estimates are consistent for varying T. In contrast to this stability, as T increases, the Static Model estimates significant positive errors and the Linear Model estimates significant negative errors. Both are trying to account for the additional polls showing large support for the Republican early in the campaign. Both alternative Models estimate much larger excess variability to compensate for the apparent misspecification of the electorate's preferences. The additional polls showing large support for the Republican candidate require that the linear trend in Model 2 slope steeply downward, which means that the Shirani-Mehr et al. (2018) model estimates that the polls likely overstate the *Democrat's* support when polls showing broad support for the *Republican* are added to the sample. Our Random Walk Model avoids this issue with its flexible, non-linear model for preferences and higher weight given to polls close to the election. Florida in 2016 shows a similar but less extreme example (right panels of Figure 4). There was a late shift in support towards candidate Trump, and polls very close to the election were fairly accurate. All models estimate 95% CIs that include zero for $T = 10$, but as more polls are added, both Static and Linear Models become increasingly confident that the polls overstated candidate Clinton's support, while our model weights the accurate polls conducted close to the election higher and avoids making that same mistake.

Next, because the Static Model is essentially a special case of the Linear Model ($\beta_r = 0$) and of the RW Model ($\gamma_r = 0$), we focus our comparison on the Linear and RW Models. Figure 5 plots the election day error and excess polling variability for all states with points colored by the election year for the Linear Model (using $T = 20$) and the RW Model (using $T = 50$). The Linear Model's T was chosen to match the implementation in Shirani-Mehr et al. 2018, while using different values of T for the RW Model does not change the results. Both models estimate large, negative errors in 2016 and 2020, indicating overstatement of Democratic support in polls, with the RW Model estimating slightly larger over estimates. The Linear Model estimates much larger τ_r terms, indicating that a poll large enough to ignore the binomial variability would have a margin of error about 2 to 3 times larger than the margin of error estimated by the RW Model. Without much flexibility for measuring how preferences evolve, the Linear Model attributes violations of the linearity assumption to merely large variability. By allowing for any kind of temporal evolution in preferences, our RW Model estimates notably smaller variance terms.

As a final highlight, across every election, we calculate the range of point estimates (posterior mean) across poll inclusion windows. Figure 6 shows the results. Our model, in blue, provides remarkably stable estimates. The range of point estimates, especially in 2004 and 2008, is narrower and usually a subset of the Linear Model's range. This indicates that not only is our model more stable across varying windows, but also it is providing similar and more precise information.

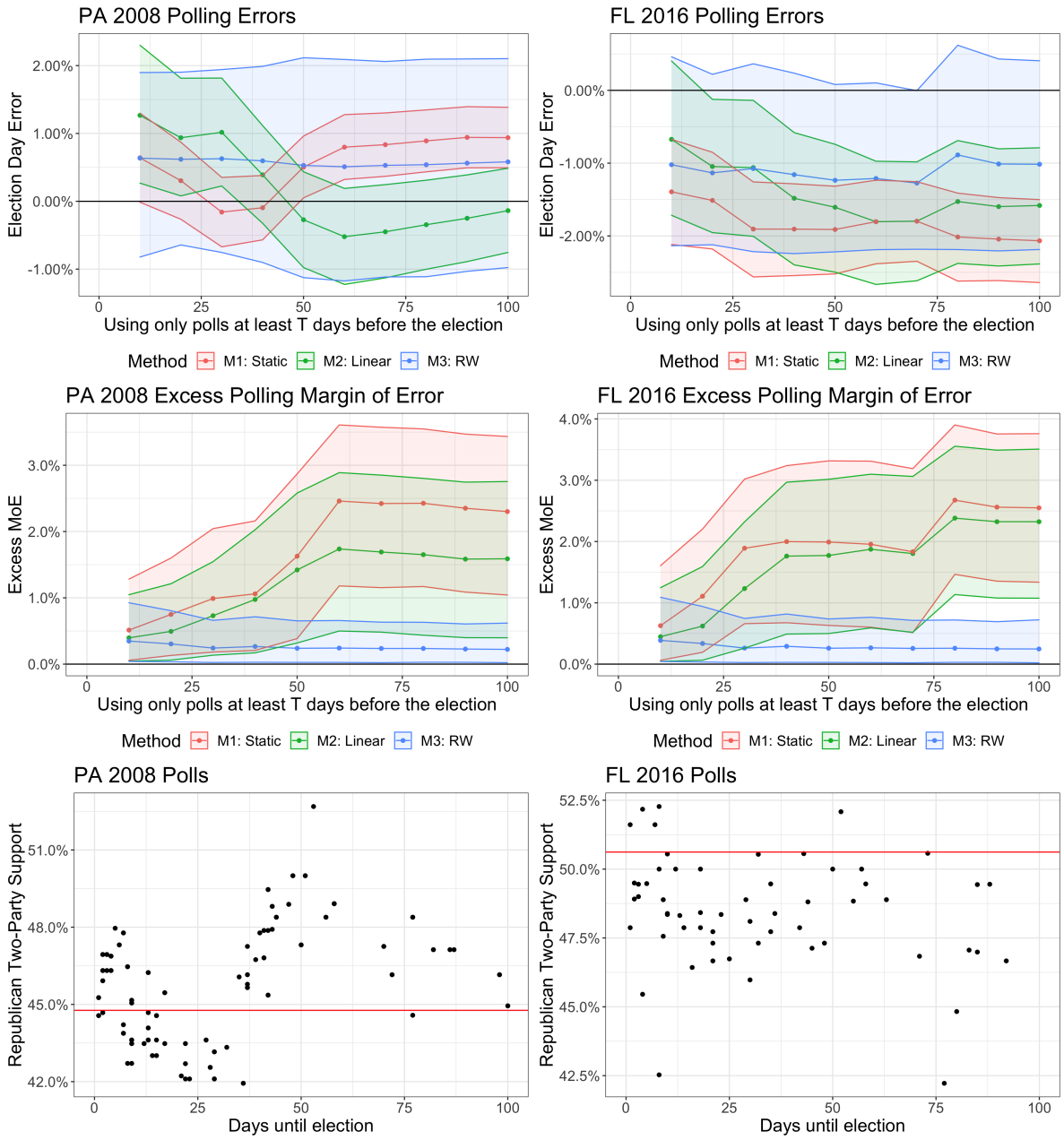

Figure 4: **Estimated Polling Error and Excess Variability by Inclusion Window for Select Elections.** Election Day Error measures the expected overestimate of Republican support for a poll conducted on election day. Shaded regions show 95% credible intervals. The bottom row shows the underlying polling data; the red line indicates the election result. Both the Static and Linear Models are sensitive to which polls are included. When their assumptions about preferences are incorrect, evidenced by the non-linearity in polls, those models increase their estimates of excess sampling variably to compensate. Our Random Walk Model, with its flexible model for preferences, does not make such erroneous estimates.

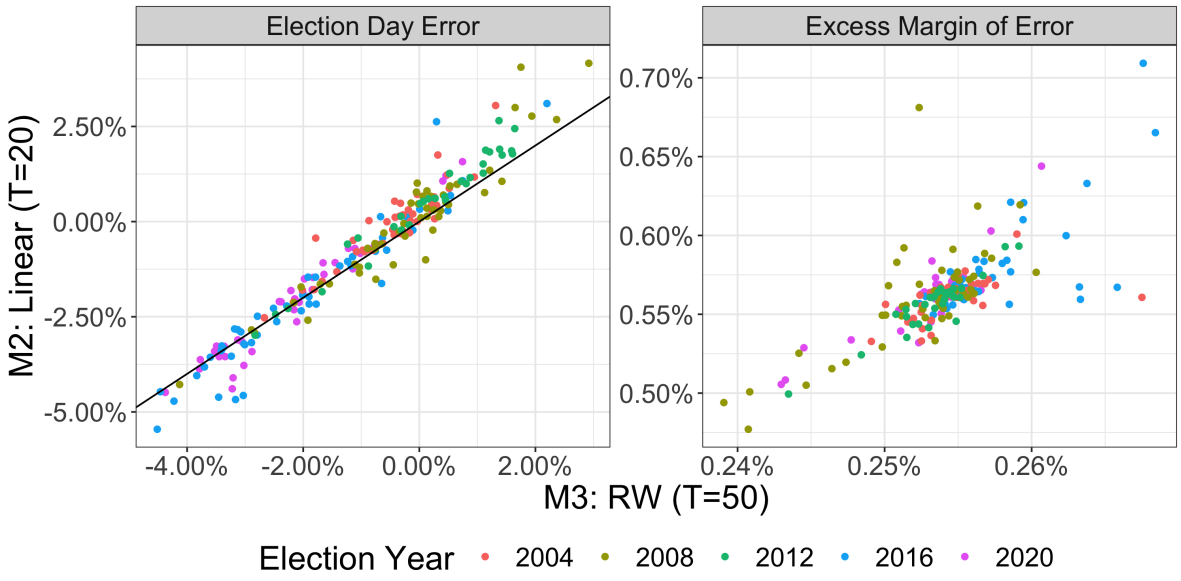

Figure 5: **Comparison of Election Day Error and Excess MoE Estimates for the Linear (M2) and Random Walk (M3) Models.** Both models estimate roughly consistent election day error, with the RW Model estimating larger overstatement of Clinton's and Biden's support in 2016 and 2020 and the Linear Model estimating larger overstatement of Romney's and McCain's support in 2008 and 2012. The Linear Model estimates that a large poll would have a margin of error about two to three times larger than the RW Model's estimate.

6 Conclusion

Pollsters have tried to adapt to the failures in past elections by changing sampling methodologies and weighting schemes, and certainly future forecasting models will pay more attention to accounting for uncertainty and error in polling data (Skelley and Rakich 2020). How one actually measures that uncertainty and error, especially as inputs to a forecast or campaign-related decisions, is an important initial undertaking. Use of simple models is appealing, and assuming constant or linear changes in preferences is in many cases a valid approach. However, over reliance on those models is cautioned against when violations of the model assumptions are not reflected in increased parameter uncertainty and the model conclusions vary based on subjective analyst decisions. The use of the more complex random walk framework allows our model to adapt to those simple assumptions only when they are justified in the data, and, when those assumptions do not hold, our model does not mislabels changes in preferences as polling error.

Our model better captures turbulent election cycles, has easily interpretable results, and most importantly measures polling errors in a way that is robust to subjective inclusion decisions about how many polls to include. When voter preferences are changing or late-breaking news stories alter election dynamics, our model is flexible enough to separate polling error from shifting preferences. This flexibility is further reflected in the much lower variability of estimates when changing how many polls are included. Other methods, with stronger assumptions about how preferences evolve, have marked shifts in estimated bias when more data are included. Moreover, with a single parameter to estimate bias, the results of our model are directly interpretable and do not vary with the actual election result, which eases the use of the model's conclusions in election forecasting and other decision-making contexts where elections results are not yet known.

As highlighted in Section 2, the potential sources of polling errors are varied and difficult to separate. Our method confirms that these additional factors do contribute to excess variability error and, for 2016 and 2020 presidential elections, notable directional errors that overstated the Democratic candidate's support. However, our model also indicates that the excess variability is much smaller than traditional estimates from models that do not separate changes in the target estimand (the electorate's preferences) from variability of the estimator (the poll result).

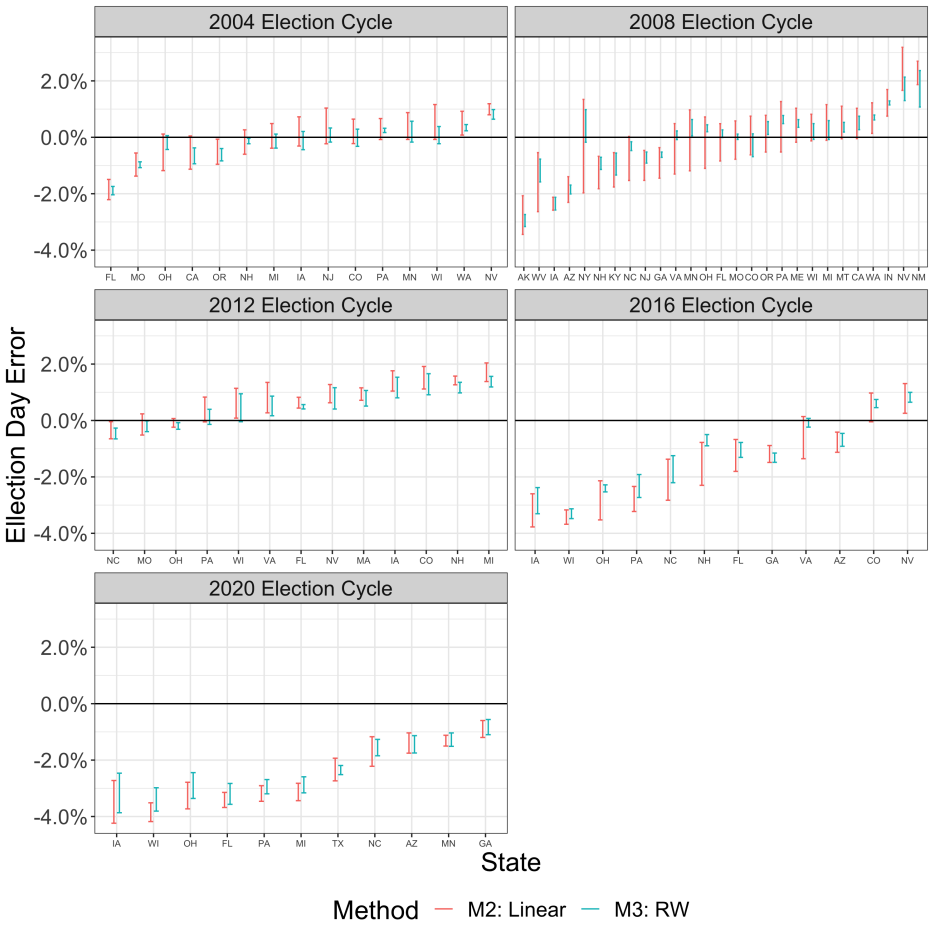

Figure 6: **Range of Polling Error Estimates.** Each bar shows the range of polling error estimates (the minimum and maximum posterior means) for the given state-year pair by model. The range is computed over 10 inclusion windows, including only polls every multiple of 10 days before the election, the same as in Figure 4. Our proposed model almost always has a narrower range and is a subset of the Linear Model's range of estimates, indicating that our model is much less sensitive to the subjective analyst decision on which polls to include. Only contests polled more than 20 times are shown

One could use our model to try and attribute polling error to poll features, such as sample frame or house effects, by separating α_r into poll-specific indicator variables or one could attribute polling error to state-level features by replacing α_r with $\mathbf{X}_r\beta_r$, where $\mathbf{X}_r$ is a vector of election-specific features such as the proportion of white voters without a college degree. This decomposition is beyond the scope of this letter as it carries substantial new complications: as the granularity of the decomposition increases, the amount of data available for any particular indicator decreases, which in turn requires more complex sharing of information across states and election years (e.g. for contemporaneous elections).

Estimating total polling error and separating bias from variance is a difficult task. With the frequency of systematic errors in recent elections, increased attention has come to this challenge. While these different models often lead to similar conclusions, ensuring that the parameters are easily interpretable and that the conclusions are not sensitive to analyst decisions, such as the inclusion window, is important. The simple interpretation of α_r in our model, as opposed to needing to compute a more complicated logistic transformation, eases the use of our model's results in other scenarios, e.g. election forecasting (which could even be done with our model by simply placing a prior on v_r), and enables assessing more complex explanations for polling errors. With a single parameter measuring error independent of the election result, researches can explore how poll or election-level features impact polling errors without ambiguity regarding

the outcome of interest.

Funding

GT and AV were partially supported by the Provost at Duke University via the Polarization Lab. AV was partially supported by the National Science Foundation (DMS-2046880).

Acknowledgement

We thank participants at the Joint Statistical Meetings and the Junior ISBA workshop for their helpful comments. We also thank D. Sunshine Hillygus and Brian Guay for helpful discussions and reviews of early drafts of this work.

Data Availability

Replication data and code for this article is available at this GitHub repository and will be placed on Dataverse upon acceptance.

References

Abramowitz, A. I. 2008. "Forecasting the 2008 presidential election with the time-for-change model." *PS: Political Science & Politics* 41 (4): 691–695.

Barnes, P. 2016. "Reality Check: Should We Give Up on Election Polling?" *BBC News* (November 16). Accessed January 25, 2022. `https://www.bbc.com/news/election-us-2016-37949527`.

Biemer, P. P. 2010. "Total survey error: Design, implementation, and evaluation." *Public opinion quarterly* 74 (5): 817–848.

Campbell, J. E., and K. A. Wink. 1990. "Trial-heat forecasts of the presidential vote." *American Politics Quarterly* 18 (3): 251–269.

Caughey, D., and C. Warshaw. 2015. "Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model." *Political Analysis* 23 (2): 197–211. doi:`10.1093/pan/mpu021`.

Chen, Y., R. Garnett, and J. M. Montgomery. 2022. "Polls, Context, and Time: A Dynamic Hierarchical Bayesian Forecasting Model for US Senate Elections." *Political Analysis:* 1–21.

Clinton, J., J. Agiesta, M. Brenan, C. Burge, M. Connelly, A. Edwards-Levy, B. Fraga, et al. 2021. "Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls." *American Association for Public Opinion Research.*

Clinton, J. D., and S. Rogers. 2013. "Robo-Polls: Taking Cues from Traditional Sources?" *PS: Political Science & Politics* 46 (2): 333–337.

Erikson, R. S., and C. Wlezien. 2008. "Leading economic indicators, the polls, and the presidential vote." *PS: Political Science & Politics* 41 (4): 703–707.

Fey, M. 1997. "Stability and coordination in Duverger's law: A formal model of preelection polls and strategic voting." *American Political science review* 91 (1): 135–147.

Gelman, A., S. Goel, D. Rivers, D. Rothschild, et al. 2016. "The mythical swing voter." *Quarterly Journal of Political Science* 11 (1): 103–130.

Groves, R. M., and L. Lyberg. 2010. "Total survey error: Past, present, and future." *Public opinion quarterly* 74 (5): 849–879.

Heidemanns, M., A. Gelman, and G. E. Morris. 2020. "An updated dynamic Bayesian forecasting model for the US presidential election." *Harvard Data Science Review* 2 (4).

Hillygus, D. S., and B. Guay. 2016. "Polling in the United States." In *The Seminar Magazine issue on 'Measuring Democracy,* vol. 684.

Hillygus, D. S. 2011. "The Evolution of Election Polling in the United States." *Public Opinion Quarterly* 75, no. 5 (December): 962–981. ISSN: 0033-362X. doi:10.1093/poq/nfr054. eprint: https://academic.oup.com/poq/article-pdf/75/5/962/5180996/nfr054.pdf. https://doi.org/10.1093/poq/nfr054.

Huang, T., and D. Shaw. 2009. "Beyond the battlegrounds? Electoral college strategies in the 2008 presidential election." *Journal of Political Marketing* 8 (4): 272–291.

Jackman, S. 2005. "Pooling the polls over an election campaign." *Australian Journal of Political Science* 40 (4): 499–517.

Jackson, N. 2018. "The rise of poll aggregation and election forecasting." In *The Oxford Handbook of polling and survey methods,* edited by L. Atkeson and R. Alvarez. Oxford: Oxford University Press.

———. 2020. "Poll-Based Election Forecasts Will Always Struggle With Uncertainty." *Center for Politics* (August 6). Accessed January 31, 2022. https://centerforpolitics.org/crystalball/articles/poll-based-election-forecasts-will-always-struggle-with-uncertainty/.

Jennings, W., and C. Wlezien. 2018. "Election polling errors across time and space." *Nature Human Behaviour* 2 (4): 276–283. doi:10.1038/s41562-018-0315-6. https://doi.org/10.1038/s41562-018-0315-6.

Kennedy, C., M. Blumenthal, S. Clement, J. D. Clinton, C. Durand, C. Franklin, K. McGeeney, et al. 2017. "An evaluation of 2016 election polls in the US." *American Association for Public Opinion Research.* Accessed January 25, 2022. https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx.

———. 2018. "An evaluation of the 2016 election polls in the United States." *Public Opinion Quarterly* 82 (1): 1–33.

Levine, D. K., and T. R. Palfrey. 2007. "The paradox of voter participation? A laboratory study." *American political science Review* 101 (1): 143–158.

Linzer, D. A. 2013. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* 108 (501): 124–134. doi:10.1080/01621459.2012.737735. eprint: https://doi.org/10.1080/01621459.2012.737735. https://doi.org/10.1080/01621459.2012.737735.

McDermott, M. L., and K. A. Frankovic. 2003. "Horserace Polling and Survey Method Effects: an analysis of the 2000 campaign." *The Public Opinion Quarterly* 67 (2): 244–264.

McFarland, S. G. 1981. "Effects of question order on survey responses." *Public Opinion Quarterly* 45 (2): 208–215.

Pickup, M., and R. Johnston. 2007. "Campaign trial heats as electoral information: evidence from the 2004 and 2006 Canadian federal elections." *Electoral Studies* 26 (2): 460–476.

———. 2008. "Campaign trial heats as election forecasts: Measurement error and bias in 2004 presidential campaign polls." *International Journal of Forecasting* 24 (2): 272–284.

Prado, R., and M. West. 2010. *Time series: modeling, computation, and inference.* Chapman / Hall/CRC.

Rothschild, D. 2009. "Forecasting elections: Comparing prediction markets, polls, and their biases." *Public Opinion Quarterly* 73 (5): 895–916.

Shirani-Mehr, H., D. Rothschild, S. Goel, and A. Gelman. 2018. "Disentangling Bias and Variance in Election Polls." *Journal of the American Statistical Association* 113 (522): 607–614. doi:10.1080/01621459.2018.1448823. eprint: https://doi.org/10.1080/01621459.2018.1448823. https://doi.org/10.1080/01621459.2018.1448823.

Skelley, G., and N. Rakich. 2020. "What Pollsters Have Changed Since 2016 — And What Still Worries Them About 2020." *FiveThirtyEight* (October). https://fivethirtyeight.com/features/what-pollsters-have-changed-since-2016-and-what-still-worries-them-about-2020/.

Smith, T. W. 1987. "That which we call welfare by any other name would smell sweeter an analysis of the impact of question wording on response patterns." *Public opinion quarterly* 51 (1): 75–83.

Stan Development Team. 2020. *RStan: the R interface to Stan.* R package version 2.21.2. http://mc-stan.org/.

Stoetzer, L. F., M. Neunhoeffer, T. Gschwend, S. Munzert, and S. Sternberg. 2019. "Forecasting elections in multiparty systems: a Bayesian approach combining polls and fundamentals." *Political Analysis* 27 (2): 255–262.

Sturgis, P., N. Baker, M. Callegaro, S. Fisher, J. Green, W. Jennings, J. Kuha, B. Lauderdale, and P. Smith. 2016. "Report of the inquiry into the 2015 British general election opinion polls." *NCRM, British Polling Council, Market Research Society.* https://eprints.soton.ac.uk/390588/1/Report_final_revised.pdf.

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

Weisberg, H. F. 2009. *The total survey error approach.* University of Chicago Press.