Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 566–582
https://doi.org/10.1093/jrsssa/qnae066
Advance access publication 25 July 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Bias and excess variance in election polling: a not-so-hidden Markov model

Graham Tierney and Alexander Volfovsky

Department of Statistical Science, Duke University, Durham, NC 27708, USA

Address for correspondence: Graham Tierney, Department of Statistical Science, Duke University, Durham, NC 27708, USA. Email: gtierney2@gmail.com

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. Analysing 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 peek 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 and 1992 through 2020 US Senate elections, concluding that previously reported estimates of bias in Presidential elections were too extreme by 10%, estimated bias in Senatorial elections was too extreme by 25%, and excess variability estimates were also too large.

Keywords: Bayesian models, election forecasting, election polls, total survey error

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 (Fey, 1997; Huang & Shaw, 2009; Levine & Palfrey, 2007). Recent high-profile polling misses both in the U.S. and the UK 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, 2018). 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 & 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 underestimating one candidate's support, and excess variance, variability above what would be implied if polls were independent random samples of the electorate. Directional error could occur if one candidate's supporters are less likely to respond to pollsters. Excess variance could occur if pollsters have different sampling methodologies that target different populations. Directional errors can be thought of as systematic errors that favour one candidate, while excess variance can be thought of as random error that does not result in bias but does make each poll a 'noisier' estimate. Estimating both of these quantities requires comparing a poll's result to the underlying value it measures. Making such comparisons is complicated by the fact that the 'ground truth' of voters'

Received: February 17, 2023. **Revised:** June 24, 2024. **Accepted:** June 25, 2024

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 expected.

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 & Wlezien, 2018) or specifying a simple (linear) model for how preferences might change over time (e.g. Shirani-Mehr et al., 2018), then estimating polling error as the difference between poll-estimated preferences and actual preferences in the election. 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, simple linear models for preferences require modelling polling errors on the logistic scale to avoid extrapolations that imply greater than 100% support for a candidate. 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, discrete-time hidden Markov model for how preferences change over time and treat the election outcome as a peek at the typically hidden, underlying Markovian process. This flexibility allows us to leverage more polling data and automatically down-weight early polls when the electorate's preferences are volatile. In essence, we model polls as noisy measurements of the electorate's preferences, which evolve over time via a random walk (RW). Then, we treat the election itself as a peek at the exact, previously hidden preferences of the electorate on election day, and use that peek to infer the directional error and excess variance of the earlier polls. The concept of such a 'peek' at the hidden process is novel because these kinds of reveals rarely, if ever, happen in other applied settings.

Our approach directly builds upon the methodology in Shirani-Mehr et al. (2018) and has three principle advantages. First, our method is much more robust to the choice of how many days of polling data are included. The hidden Markov model for the electorate's preferences naturally discounts early polls if preferences are highly volatile and leverages them if preferences are stable. Figure 1 shows this phenomenon, highlighting in purple states where the poll-favoured candidate actually changes between the Democrat and the Republican depending on how many days of polling are included in the model. The top row shows that the Shirani-Mehr et al. (2018) model is much more sensitive than the bottom row with our model's results. Second, we avoid conflating changes in preferences with polling error. We model how preferences change over time, so if late-breaking news stories shift support from one candidate to another, we avoid conflating that shift with polling error. Third, our model is much more interpretable. We use a flexible RW rather than a linear model for preferences, so we can directly model error without a logistic transformation complicating interpretation as in Shirani-Mehr et al. (2018). A logistic transformation makes the key quantity, the amount of error, depend on the final election result.

We estimate that polls did not systematically over- or under-state either party across state-level contests, while they did overstate the Democrats' support by approximately 2 percentage points in 2016 and 2020 Presidential elections. In contrast to directly comparing the poll and election result or using a linear trend adjustment (Shirani-Mehr et al., 2018), we estimate less extreme errors by a factor of 10% to 25% 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.

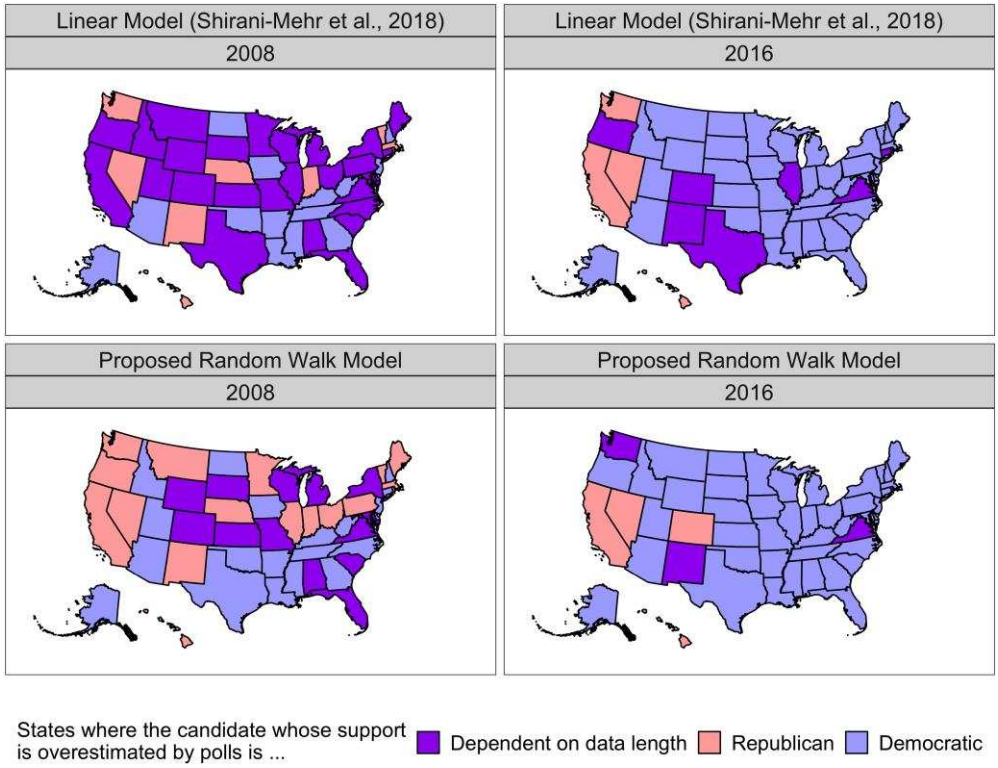

Figure 1. Shifting results by inclusion window in presidential elections. Figure shows for the 2008 and 2016 election cycles, the two most-polled years in our data, states where polling error estimated by the linear model of Shirani-Mehr et al. (2018) and our proposed model (Section 4) *changed signs*. States where the directional error could favour *either* the Democrat or the Republican candidate depending on how many days of polling are used in estimation are shaded purple. Our model (bottom row) is much more consistent in which candidate is favoured because it adaptively learns and down-weights polls conducted long before the election.

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.

2.1 Polling accuracy

Pre-election or 'horserace' polls have a long history in the U.S. 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 (Biemer, 2010; Groves & Lyberg, 2010; Weisberg, 2009). 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. While we do not measure these sources explicitly, rather estimating aggregate errors

for all polls of each race, we discuss possible extensions to our model that could quantify these sources of error in Section 6.

After the high-profile polling misses in the state-level results of the 2016 and 2020 US Presidential elections, along with misses in the 2015 UK general election and 2016 Brexit referendum, 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 (Bon et al., 2019; Clinton et al., 2021; Kennedy et al., 2017; Sturgis et al., 2016, 2018).

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 & 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. More recently, Bon et al. (2019) used this model to decompose bias due to undecided voters from other sources with special attention to the 2016 US Presidential election, which had an unusually high number of undecided voters close to the election. We detail the Shirani-Mehr et al. (2018) 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). Even earlier, researchers developed methods to combine polls and public opinion surveys to separate real changes in preferences from survey error and identify the impact of campaign events (Erikson & Wlezien, 1999; Green et al., 1999; Wlezien & Erikson, 2002). We focus this section on only the specific class of forecasting models that we leverage in our method to estimate polling errors, highlighting key innovations and areas where our application simplifies certain assumptions. Note that these models all treat the election outcome as uncertain and use polls (along with other data) to predict the election outcome. We will use the same modelling principles but treat the election outcome as known and estimate how accurate or inaccurate the polls were. Pasek (2015) provides a detailed review of alternative methods for election forecasting and poll aggregation beyond the hidden Markov approach we focus on here.

Linzer (2013) developed a hidden Markov model for US Presidential election forecasting where underlying state-level preferences evolve over time following a RW. 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, was developed for general public opinion tracking in Green et al. (1999), and has been expanded and applied by many forecasting models, including multi-party systems (Stoetzer et al., 2019; Walther, 2015), and the Economist's 2020 forecast, which made additional adjustments for correlated shifts in state-level trends and polling errors (Heidemanns et al., 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 et al., 2022). The general modelling framework whereby surveys measure latent preferences has been applied beyond election polls (e.g. Caughey & 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 & Wink, 1990; Erikson & Wlezien, 2008). We, however, will condition on the election outcome to estimate polling errors, removing the need for forecasting and applying fundamentals approaches altogether. This assumption is appropriate in our case, as we are interested in a retrospective analysis of polling error and so can assume that the election has already occurred. While not the primary purpose of our development, in Section 6, we describe how forecasting with our model is feasible by placing a prior on the election result. 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.

Recent work has also attempted to use non-representative polls via multilevel regression and post-stratification (Gelman et al., 2016; Hanretty, 2020; Kiewiet de Jonge et al., 2018; Wang et al., 2015). This is in essence a re-weighting procedure where candidate preferences are regressed on demographic features of poll respondents, then projected to the target population (the entire electorate or likely voters) to adjust for over- or under-sampling of certain groups in the poll. This approach focuses mostly on adjusting results poll-by-poll rather than the poll aggregation that our model focuses on. It also requires detailed respondent-level data that are not as frequently reported (and not reported in the data we use).

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 & 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. We also rely on the Senate polls collected in Chen et al. (2022), who use them to train an election forecasting model. The Senate polls cover elections from 1992 to 2020. We limit our data to only polls conducted in the 100 days before any election because that is the smallest window we have for the earlier 2004 to 2012 elections. Moreover, given results already shown in Figure 1 and later detailed in Section 5, results are extremely unlikely to differ when using longer windows. All data and code are available at this GitHub repository.

The availability of polling data varies across elections and states. Election 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. Senatorial elections are polled less frequently than Presidential elections, but there has been a notable increase in the number of polls over time. Figure 3 shows the number of polls conducted at most 100 days before the election by state and year for Presidential contests. The 2008 election cycle again stands out, as do many swing states such as Ohio, Pennsylvania, and Florida. We do not show a similar map for Senatorial elections because of the timing irregularity of Senate contests; not every state has a Senate race in every cycle.

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

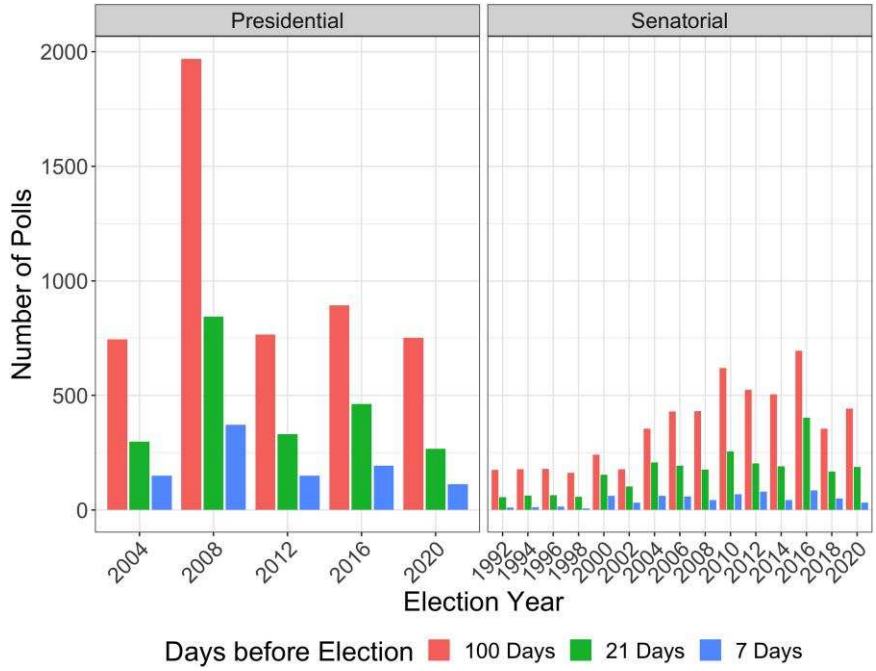

Figure 2. Number of polls by varying time window. Columns show the number of polls conducted each election cycle; colours indicate the cut-off time. Polls are conducted more frequently towards the end of the election cycle; approximately 19% and 43% of polls conducted in the final 100 days are conducted in the final 7 and 21 days, respectively.

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 cut-off 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*, true underlying preferences measured by 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 cut-off 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 linearly 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].

M3: 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 RW: $\theta_{r,t+1} \sim N(\theta_{rt}, \gamma_r^2)$. 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 RW 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]. A detailed discussion of this model is presented in the next section along with a clarifying plate diagram in Figure 4.

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 modelling decision is

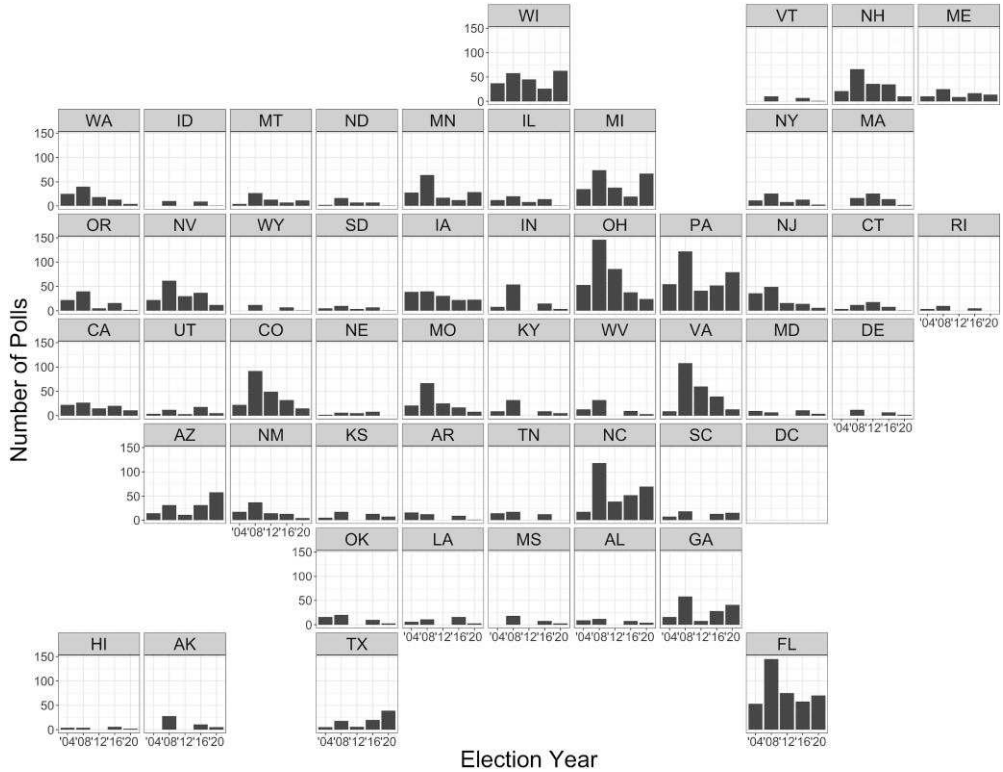

Figure 3. Number of presidential polls by state and year. Columns show the total number of polls conducted within 100 days of the election for each election cycle in each state. Swing states are polled much more frequently than non-swing states, and many states are not polled at all in some election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

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 $\text{Var}(y_i \mid p_i) = p_i(1-p_i)/n_i$. 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 RW 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, we formally state the model.

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

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

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

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

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

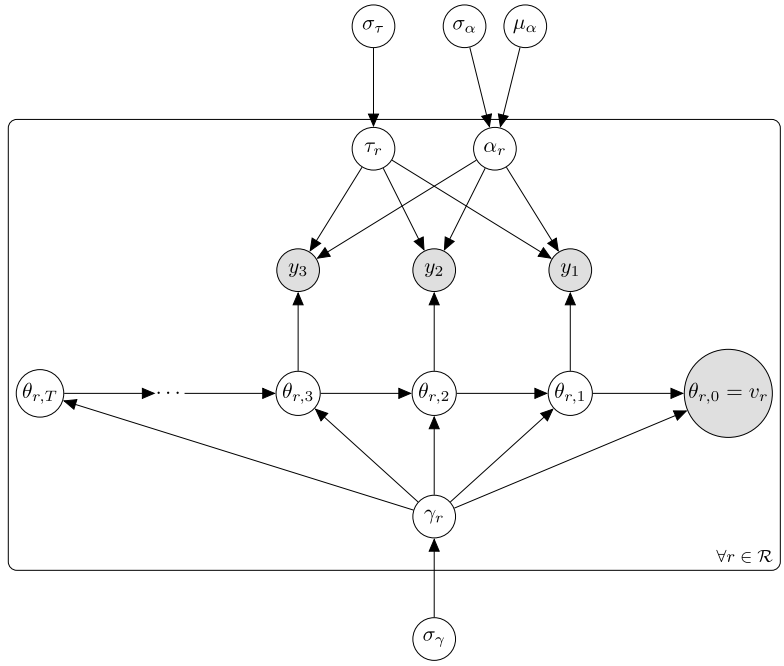

Figure 4. Diagram of RW model. This diagram represents the dependencies in equations (1)–(4) above. θ_{rt} represents the electorate's preferences in election r at time t. Any number of polls can be conducted on each day (including zero). Polls are informed by preferences θ_{rt}, directional error α_r, and excess variance τ_r^2 for their corresponding election and day. Information is shared across elections via the hierarchical model on α_r, τ_r, and γ_r governed by hyperparameters denoted with subscripts. Note that p terms have been omitted because $p_{rt} = \theta_{rt} + \alpha_r$ almost always.

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. 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. 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 with slight modifications to account for the lack of a logistic transformation: $\mu_\alpha \sim N(0, 0.05^2)$, $\sigma_\alpha \sim N_+(0, 0.2^2)$, $\sigma_\tau \sim N(0, 0.05^2)$. For 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. Figure 4 shows a plate diagram of the model illustrating the dependencies between parameters. Note that observed nodes are shaded, including the final, election-day preferences $\theta_{r,0}$. The diagram also reveals that integrating out the unobserved RW parameters will allow us to express election day error α_r and excess variance τ_r as functions of the polling data and election outcome. This derivation is outlined in Section 4.2 below. Inference proceeds directly via equations (1)–(4) which describe the observed data likelihood and RW evolution; this effectively integrates over days without polls.

The model is estimated in Stan using Hamiltonian Monte Carlo (Stan Development Team, 2020). We also improve convergence by reparameterizing 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 centred parameterization described in Prado and West (2010).

4.2 Use of observed errors $y_i - v_r$

In this section, we discuss how each model estimates α_r as a function of observed polling errors $y_i - v_r$. We will show that our model can be thought of as a temporally weighted average of observed polling errors, whereas the static model is an equally weighted average that ignores temporal information and the linear model does not have any such representation. 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.

Static model. 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:

$$\alpha_r \mid \mathbf{Y}, M_1 \sim N\left(\frac{n_r/\tau_r^2}{n_r/\tau_r^2 + \sigma_\alpha^{-2}} \frac{\sum_i (y_i - v_r)}{n_r} + \frac{\sigma_\alpha^{-2}}{n_r/\tau_r^2 + \sigma_\alpha^{-2}} \mu_\alpha, \left(n_r/\tau_r^2 + \sigma_\alpha^{-2}\right)^{-1}\right), \tag{5}$$

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 the 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.

Linear model. 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 $y_i - v_r$. Moreover, because of the linear assumption, additional polls showing large support for candidate A when the election result was close to 50–50, can actually cause the linear model to estimate that polls favour candidate B. The linear trend needs to become steeper to fit the early polls, which shifts the intercept α_r in the opposite direction of what the new data indicate. We demonstrate that this phenomenon does indeed happen in Section 5.

RW model. The proposed RW 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:

$$\alpha_r \mid y_i, \tau_r, \gamma_r, M_3 \sim N\big(w_i(y_i - v_r) + (1 - w_i)\mu_\alpha, \lambda_i^{-1}\big), \tag{6}$$

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. As close inspection of Figure 4 will reveal, 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: $\mathbf{Y} \sim N(\nu\mathbf{1} + \alpha\mathbf{1}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \tau^2\mathbf{I} + \gamma^2\mathbf{T}$ and $T_{ij} = \min(t_i, t_j)$. Under $p(\alpha) \propto 1$, we have

$$p(\alpha \mid \mathbf{Y}) \propto \exp\left(-\frac{1}{2}(\mathbf{Y} - \nu\mathbf{1} - \alpha\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \nu\mathbf{1} - \alpha\mathbf{1})\right) \Rightarrow \tag{7}$$

$$\alpha \mid \mathbf{Y} \sim N\left((\mathbf{Y} - \nu\mathbf{1})'\boldsymbol{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}), \left(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\right)^{-1}\right) \tag{8}$$

That is, the posterior mean of α is a weighted average of the observed polling errors, $y_i - \upsilon_r$ 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.

Thus, we can already see that the three advantages of our model noted in Section 1 are clear from the construction. First, our model is more consistent because it learns the weighting scheme described above to down-weight early polls of the race. Second, with the flexible RW structure instead of a linear time trend, our model avoids conflating changes in preferences with polling error. And finally, our model is much more interpretable. Without the logistic transformations required by the linear time trends, one can directly interpret α_r and see how the estimate is a weighted average of observed polling errors.

5 Comparisons

We compare the models based on election day error (α_r for the static and RW models and $\text{logit}^{-1}(\text{logit}(\upsilon_r) + \alpha_r) - \upsilon_r$ 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). Positive values of election day error indicate the Republican candidate's support is overstated by polls. We estimate each model 60 times increasing the time cut-off T by 1 day increments between 1 and 50 days before the election and 5 day increments between 50 and 100 days before the election, i.e. $T \in \{1, 2, \ldots, 49, 50, 55, 60, \ldots, 95, 100\}$.

Figure 5 highlights two Presidential elections where the three models give different results. In the 2008 election in Pennsylvania (left panels of Figure 5), 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 RW 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 5). 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. Recall that in Figure 1 we showed that the linear Model 2 was indeed less robust to the length of the data across multiple states and election cycles beyond the two examples shown here. Indeed, the linear model changes which candidate's support is overestimated in the polls when different date cut-offs T are chosen in 30% of Presidential contests (67/221). Our RW model has this feature only 17% of the time.

Next, to further highlight that the above are not isolated examples, we compare the models' election day error and excess variance estimates across all contests in our data using $T = 21$ to match the implementation in Shirani-Mehr et al. (2018). Because the Static Model is essentially a special case of both the linear model ($\beta_r = 0$) and of the RW model ($\gamma_r = 0$), we focus our

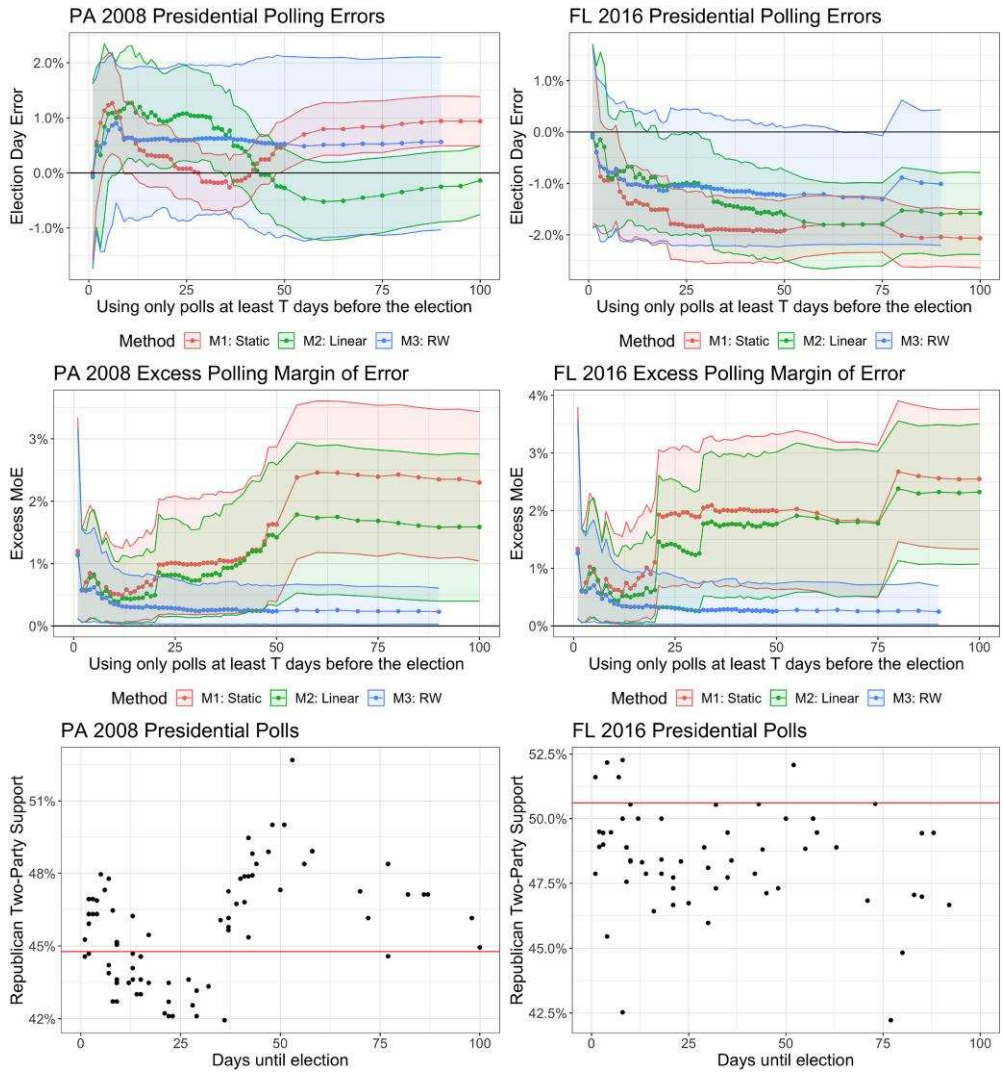

Figure 5. Estimated polling error and excess variability by inclusion window for select elections. Election day error measures the expected overestimate of Republican support for a poll conducted on election day. Shaded regions show 95% credible intervals. The bottom row shows the underlying polling 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 RW model, with its flexible model for preferences, does not make such erroneous estimates.

comparison on the linear and RW models. Figure 6 plots the election day error (top row) and excess margin of error (bottom row) for Presidential (left column) and Senatorial (right column) elections with colours indicating the election year.

The Linear Model consistently estimates more extreme election day error than the RW model across both kinds of elections. Both models indicate that polls overestimated Secretary Clinton and President Biden's support in 2016 and 2020 (negative errors) but the linear model's estimates are larger in magnitude. A similar pattern holds for contests with positive errors when polls overstate the Republican's support: the linear model estimates larger errors than our RW model. This is directly attributable to the model assumptions. *Any* non-linear change in preferences in the 3 weeks before the election must be directional polling error in the Linear Model by assumption, whereas our more flexible approach avoids that mistake.

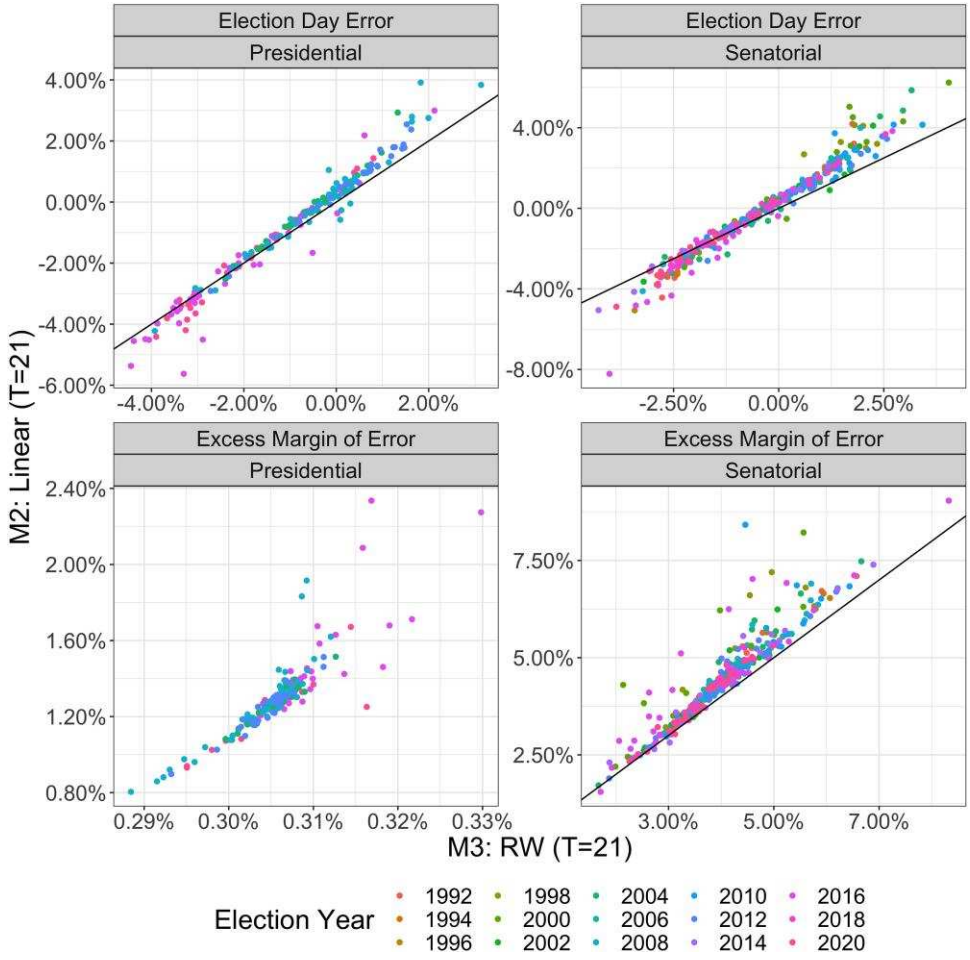

Figure 6. Comparison of election day error and excess MoE estimates for the linear (M2) and RW (M3) models. Each point compares either the election day error (top row) or excess margin of error (bottom row) for Presidential (left column) or Senatorial (right column) elections. All polls conducted within 21 days of the election are used, matching the implementation in Shirani-Mehr et al. (2018). The Linear Model estimates more extreme biases and larger margins of error than the more flexible RW model.

The Linear Model also estimates much larger τ_r terms, especially in Presidential elections where the unity line cannot even be shown on the plot. The linear model indicates that a Presidential poll large enough to ignore binomial sampling variability should still report a margin of error about 2 to 3 times larger than the margin of error estimated by the RW model. In Senatorial elections, both models estimate much larger excess margin of errors than in Presidential elections. The linear estimates are almost universally larger but the difference is not as extreme. Without much flexibility for measuring how preferences evolve, the linear model attributes violations of the linearity assumption to excess polling variability. By allowing for any kind of temporal evolution in preferences, our RW model estimates notably smaller variance terms.

It is important to note that the election day error result (that the Linear Model estimates more extreme errors) is not an artefact of the Shirani-Mehr et al. (2018) $T = 21$ day cut-off time, state-level dynamics, or cycle-specific features. Table 1 shows results from regressing the Linear Model's expected election day error on our RW model's estimate with varying fixed effects for state, election year, and cut-off time. We include estimates for all elections and all cut-off times between 14 and 28 days before the election. The Linear errors are consistently about 10% more extreme and about half of a percentage point larger than our model's estimates. Regression coefficients are all

Table 1. Regression analysis of presidential polling errors

	(1)	(2)	(3)	(4)
Random walk error	1.108***	1.108***	1.122***	1.129***
	(0.004)	(0.004)	(0.003)	(0.003)
Constant	0.568***	0.459***	0.399***	0.283***
	(0.041)	(0.038)	(0.037)	(0.006)
State	Yes	Yes	Yes	No
Election year	Yes	Yes	No	No
Cut-off time	Yes	No	No	No
Observations	3,186	3,186	3,186	3,186
R^2	0.983	0.983	0.982	0.977
Adjusted R^2	0.983	0.982	0.982	0.977
Residual std. error	0.237	0.242	0.245	0.277
F statistic	2,697.331***	3,275.191***	3,424.373***	134,007.400***

Note. $^*p < .05$; $^{**}p < .01$; $^{***}p < .001$. Each column reports results from regressing the linear model's estimated error on the RW model's error for the same state, year, and cut-off time. The first column includes fixed effects for all three identifiers and subsequent columns remove them. Data are all estimated errors for Presidential elections with cut-off times between 14 and 28 days before the election.

Table 2. Regression analysis of senatorial polling errors

	(1)	(2)	(3)	(4)
Random walk error	1.248***	1.248***	1.256***	1.260***
	(0.005)	(0.005)	(0.005)	(0.005)
Constant	0.508***	0.405***	0.497***	0.433***
	(0.081)	(0.080)	(0.079)	(0.007)
State	Yes	Yes	Yes	No
Election year	Yes	Yes	No	No
Cut-off time	Yes	No	No	No
Observations	2,592	2,592	2,592	2,592
R^2	0.977	0.975	0.973	0.965
Adjusted R^2	0.976	0.975	0.973	0.965
Residual std. error	0.288	0.296	0.306	0.348
F statistic	1,500.116***	1,779.242***	1,861.689***	71,411.690***

Note. $^*p < .05$; $^{**}p < .01$; $^{***}p < .001$. Each column reports results from regressing the linear model's estimated error on the RW model's error for the same state, year, and cut-off time. The first column includes fixed effects for all three identifiers and subsequent columns remove them. Data are all estimated errors for Senatorial elections with cut-off times between 14 and 28 days before the election.

significantly different from 1, which would imply equality, with $p < .001$. Table 2 reports the same analysis for Senatorial elections. While the constant is approximately the same, Linear errors are about half a percentage point larger, the RW coefficient is much larger. Linear errors are about 25% more extreme even after controlling for state, year, and cut-off fixed effects. Again, all coefficients in every specification are significantly different from 1 with $p < .001$.

Finally, we compare the variability of the error estimates across different cut-off times T. Our RW Model automatically discounts the information from early polls and changes its estimates of error only when fluctuations in preferences are relatively constant (γ_r is small). The Linear

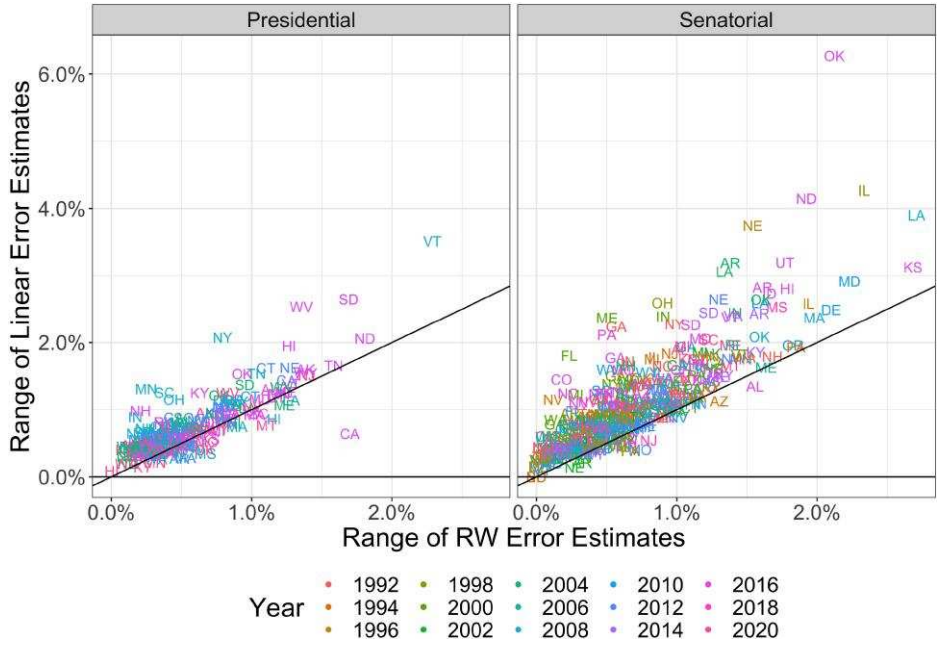

Figure 7. Comparison of the range of polling error estimates from 14 to 28 days before the election. Each point is the largest minus the smallest error estimate for the linear model (*y*-axis) and RW model (*x*-axis) across the 15 potential cut-off choices between 2 and 4 weeks prior to the election. Our proposed model almost always has a narrower range than 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.

Model weights all polls equally after removing the linear trend, so it will be more sensitive to the subjective decision of how many polls to include. We compute for each election and model the range of error estimates for cutoffs between 14 and 28 days. Figure 7 plots the linear and RW error ranges split by election type, coloured by year, and labelled with the two-letter state abbreviation. We observe that consistently the Linear model estimates are more variable, especially for Senate elections: 75% of Presidential and 82% of Senatorial elections have wider ranges of error estimates when using the linear model instead of our RW model.

6 Discussion

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 uncertainty and error in polling data (Skelley & Rakich, 2020). How one actually measures that uncertainty and error, especially for use as inputs to forecasts 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 RW 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 mislabel changes in preferences as polling error.

Our model better captures turbulent election cycles, has easily interpretable results, and 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. Simple models will attribute these changes to directional error or excess variance, reflected in the Linear Models estimates that are 10% and 25% more extreme than our model's estimates. 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 election 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 and errors are less extreme 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).

One limitation of what we have presented is that we do not identify sources of error, estimating a single α_r and τ_r per election that measures the bias and excess variance of a typical poll of that race. To attribute polling error to poll features, such as sample frame or house effects, our model could be expanded by separating α_r into poll-specific indicator variables, or one could even 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. Implementing this extension could improve generalizability. If polls shift from over the phone to online, typical polling errors in future elections could shift as well and accounting for mode of contact could improve the polling error prediction. This decomposition is beyond the scope of this paper as it carries substantial new complications: as the granularity of the decomposition increases, the amount of data available for any particular feature decreases, which in turn requires more complex sharing of information across states and election years (e.g. for contemporaneous elections).

Another potential extension of our model would be to measure third party and undecided voters. The latent space could be expanded to include additional components corresponding to Democratic, Republican, third party, and undecided voters. The reveal on election day shows third party preferences as well, and one could set the number of undecided voters to zero. While conceptually feasible, modelling the smaller proportions close or equal to zero with a RW is more challenging because the boundary is more relevant, and not all polls report data on third party and undecided voters. With more granular data this could be an important extension, especially for undecided voters because it would relax the implicit assumption that undecided voters split proportional to the candidate's current levels of support.

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 ν_r), and enables assessing more complex explanations for polling errors. With a single parameter measuring error independent of the election result, researchers can explore how poll or election-level features impact polling errors without ambiguity regarding the outcome of interest.

Acknowledgments

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.

Conflict of interest: We have no conflicts of interest to disclose for this work.

Funding

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

Data availability

Data and code are available in this GitHub repository: www.github.com/g-tierney/polling_errors_replication.

References

Abramowitz A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science & Politics*, *41*(4), 691–695. https://doi.org/10.1017/S1049096508081249

Barnes P. (2016, November 16). Reality check: Should we give up on election polling? *BBC News*. https://www.bbc.com/news/election-us-2016-37949527.Date accessed January 25, 2022.

Biemer P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Bon J. J., Ballard T., & Baffour B. (2019). Polling bias and undecided voter allocations: US presidential elections, 2004–2016. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 467–493. https://doi.org/10.1111/rssa.12414

Campbell J. E., & Wink K. A. (1990). Trial-heat forecasts of the presidential vote. *American Politics Quarterly*, *18*(3), 251–269. https://doi.org/10.1177/1532673X9001800301

Caughey D., & Warshaw C. (2015). Dynamic estimation of latent opinion using a hierarchical group-level IRT model. *Political Analysis*, *23*(2), 197–211. https://doi.org/10.1093/pan/mpu021

Chen Y., Garnett R., & Montgomery J. M. (2022). Polls, context, and time: A dynamic hierarchical Bayesian forecasting model for US senate elections. *Political Analysis*, *31*(1), 113–133. https://doi.org/10.1017/pan.2021.42

Clinton J., Agiesta J., Brenan M., Burge C., Connelly M., Edwards-Levy A., Fraga B., Guskin E., Hillygus D. S., Jackson C., Jones J., Keeter S., Khanna K., Lapinski J., Saad L., Shaw D., Smith A., Wilson D., & Wlezien C. (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., & Rogers S. (2013). Robo-polls: Taking cues from traditional sources? *PS: Political Science & Politics*, *46*(2), 333–337. https://doi.org/10.1017/S1049096513000012

Erikson R. S., & Wlezien C. (1999). Presidential polls as a time series: The case of 1996. *Public Opinion Quarterly*, *63*(2), 163–177. https://doi.org/10.1086/297709

Erikson R. S., & Wlezien C. (2008). Leading economic indicators, the polls, and the presidential vote. *PS: Political Science & Politics*, *41*(4), 703–707. https://doi.org/10.1017/S1049096508081237

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. https://doi.org/10.2307/2952264

Gelman A., Goel S., Rivers D., & Rothschild D. (2016). The mythical swing voter. *Quarterly Journal of Political Science*, *11*(1), 103–130. https://doi.org/10.1561/100.00015031

Green D. P., Gerber A. S., & De Boef S. L. (1999). Tracking opinion over time: A method for reducing sampling error. *Public Opinion Quarterly*, *63*(2),178–192. https://doi.org/10.1086/297710

Groves R. M., & Lyberg L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hanretty C. (2020). An introduction to multilevel regression and post-stratification for estimating constituency opinion. *Political Studies Review*, *18*(4), 630–645. https://doi.org/10.1177/1478929919864773

Heidemanns M., Gelman A., & Morris G. E. (2020). An updated dynamic Bayesian forecasting model for the US presidential election. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92

Hillygus D. S. (2011). The evolution of election polling in the United States. *Public Opinion Quarterly*, *75*(5), 962–981. https://doi.org/10.1093/poq/nfr054

Hillygus D. S., & Guay B. (2016). Polling in the united states. *Seminar Magazine*, *684*, 51–55.

Huang T., & Shaw D. (2009). Beyond the battlegrounds? electoral college strategies in the 2008 presidential election. *Journal of Political Marketing*, *8*(4), 272–291. https://doi.org/10.1080/15377850903263771

Jackman S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science*, *40*(4), 499–517. https://doi.org/10.1080/10361140500302472

Jackson N. (2018). The rise of poll aggregation and election forecasting. In L. Atkeson, & R. Alvarez (Eds.), *The Oxford Handbook of polling and survey methods*. Oxford University Press.

Jackson N. (2020, August 6). Poll-based election forecasts will always struggle with uncertainty. *Center for Politics*. https://centerforpolitics.org/crystalball/articles/poll-based-election-forecasts-will-always-struggle-with-uncertainty/. Date accessed January 31, 2022.

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

Kennedy C., Blumenthal M., Clement S., Clinton J. D., Durand C., Franklin C., McGeeney K., Miringoff L., Olson K., Rivers D., Saad L., Witt G. E., & Wlezien C. (2017). An evaluation of 2016 election polls in the US. American Association for Public Opinion Research. https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx.Date accessed January 25, 2022.

Kennedy C., Blumenthal M., Clement S., Clinton J. D., Durand C., Franklin C., McGeeney K., Miringoff L., Olson K., Rivers D., Saad L., Witt G. E., & Wlezien C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, *82*(1), 1–33. https://doi.org/10.1093/poq/nfx047

Kiewiet de Jonge C. P., Langer G., & Sinozich S. (2018). Predicting state presidential election results using national tracking polls and multilevel regression with poststratification (MRP). *Public Opinion Quarterly*, *82*(3), 419–446. https://doi.org/10.1093/poq/nfy023

Levine D. K., & Palfrey T. R. (2007). The paradox of voter participation? a laboratory study. *American Political Science Review*, *101*(1), 143–158. https://doi.org/10.1017/S0003055407070013

Linzer D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, *108*(501), 124–134. https://doi.org/10.1080/01621459.2012.737735

McDermott M. L., & Frankovic K. A. (2003). Horserace polling and survey method effects: An analysis of the 2000 campaign. *The Public Opinion Quarterly*, *67*(2), 244–264. https://doi.org/10.1086/374574

McFarland S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *45*(2), 208–215. https://doi.org/10.1086/268651

Pasek J. (2015). Predicting elections: Considering tools to pool the polls. *Public Opinion Quarterly*, *79*(2), 594–619. https://doi.org/10.1093/poq/nfu060

Pickup M., & Johnston R. (2007). Campaign trial heats as electoral information: Evidence from the 2004 and 2006 Canadian federal elections. *Electoral Studies*, *26*(2), 460–476. https://doi.org/10.1016/j.electstud.2007.03.001

Pickup M., & Johnston R. (2008). Campaign trial heats as election forecasts: Measurement error and bias in 2004 presidential campaign polls. *International Journal of Forecasting*, *24*(2), 272–284. https://doi.org/10.1016/j.ijforecast.2008.02.007

Prado R., & West M. (2010). *Time series: Modeling, computation, and inference*. Chapman and Hall/CRC.

Rothschild D. (2009). Forecasting elections: Comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, *73*(5), 895–916. https://doi.org/10.1093/poq/nfp082

Shirani-Mehr H., Rothschild D., Goel S., & Gelman A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*(522), 607–614. https://doi.org/10.1080/01621459.2018.1448823

Skelley G., & Rakich N. (2020, October). What pollsters have changed since 2016—and what still worries them about 2020. FiveThirtyEight. 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. https://doi.org/10.1086/269015

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mcstan.org/.

Stoetzer L. F., Neunhoeffer M., Gschwend T., Munzert S., & Sternberg S. (2019). Forecasting elections in multiparty systems: A Bayesian approach combining polls and fundamentals. *Political Analysis*, *27*(2), 255–262. https://doi.org/10.1017/pan.2018.49

Sturgis P., Baker N., Callegaro M., Fisher S., Green J., Jennings W., Kuha J., Lauderdale B., & Smith P. (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.

Sturgis P., Kuha J., Baker N., Callegaro M., Fisher S., Green J., Jennings W., Lauderdale B. E., & Smith P. (2018). An assessment of the causes of the errors in the 2015 UK general election opinion polls. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(3), 757–781. https://doi.org/10.1111/rssa.12329

Walther D. (2015). Picking the winner (s): Forecasting elections in multiparty systems. *Electoral Studies*, *40*, 1–13. https://doi.org/10.1016/j.electstud.2015.06.003

Wang W., Rothschild D., Goel S., & Gelman A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991. https://doi.org/10.1016/j.ijforecast.2014.06.001

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

Wlezien C., & Erikson R. S. (2002). The timeline of presidential election campaigns. *The Journal of Politics*, *64*(4), 969–993. https://doi.org/10.1111/1468-2508.00159