Supplementary Appendices for: Supplementary Appendix for: If a Statistical Model Predicts That Common Events Should Occur Only Once in 10,000 Elections, Maybe it's the Wrong Model

Danny Ebanks* Jonathan N. Katz† Gary King‡

May 10, 2023

*Ph.D. Candidate, California Institute of Technology; DEbanks@Caltech.edu, DannyEbanks.com.

†Kay Sugahara Professor of Social Sciences and Statistics, California Institute of Technology; JKatz.Caltech.edu, JKatz@Caltech.edu

‡Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University; GaryKing.org, King@Harvard.edu.

Contents

1 Comparison of Nominal Confidence Interval Lengths

To quantify the magnitude of uncertainty differences between the Normal and LogisTiCC models for district- and legislature-level statistics, we compute the ratios of the credible interval (CI) widths from these two models. To compute the ratio of CI widths for district-level results, we take each of the elections for which we make a prediction and compute the width of the 95% credibility interval for both the Normal and LogisTiCC models. We then calculate the ratio of the widths of the LogisTiCC CI's to the Normal. To compute the ratio of the credibility intervals for the legislative median, we compute a 95% credibility interval for the median seat in the House for each year under each model, again out-of-sample. We take the ratio for each of the 27 years for which we make a prediction, and report the density of these ratios.

Figure 1 reports distributions of these ratios, with summaries in Table 1. The table shows that, at the individual level, the LogisTiCC forecast credible intervals are only 42 percent larger than those of Gelman-King model on average, with a mode at 25 percent, which we can see from the figure. At the same time, because of the correlations between different districts represented in the LogisTiCC, its CIs for the legislative median are 500 percent larger, on average. Given the results in Figures 1–3, it is clear that these larger CIs are needed for accurate calibration due to dependence across districts.

	Mean	Standard Deviation
District Level Results	1.42	0.246
Legislature Level Results	5.06	1.19

Table 1: Numerical summaries of Figure 1

2 Ablation Studies

We make four modeling innovations to achieve generatively accurate model predictions: a national trend, coefficient stability, local uniqueness, and electoral surprises. In this section, we conduct "ablation studies," where each model component is sequentially removed to show how the model degrades. The conclusion of this section is that all model components are essential to achieve the performance we report.

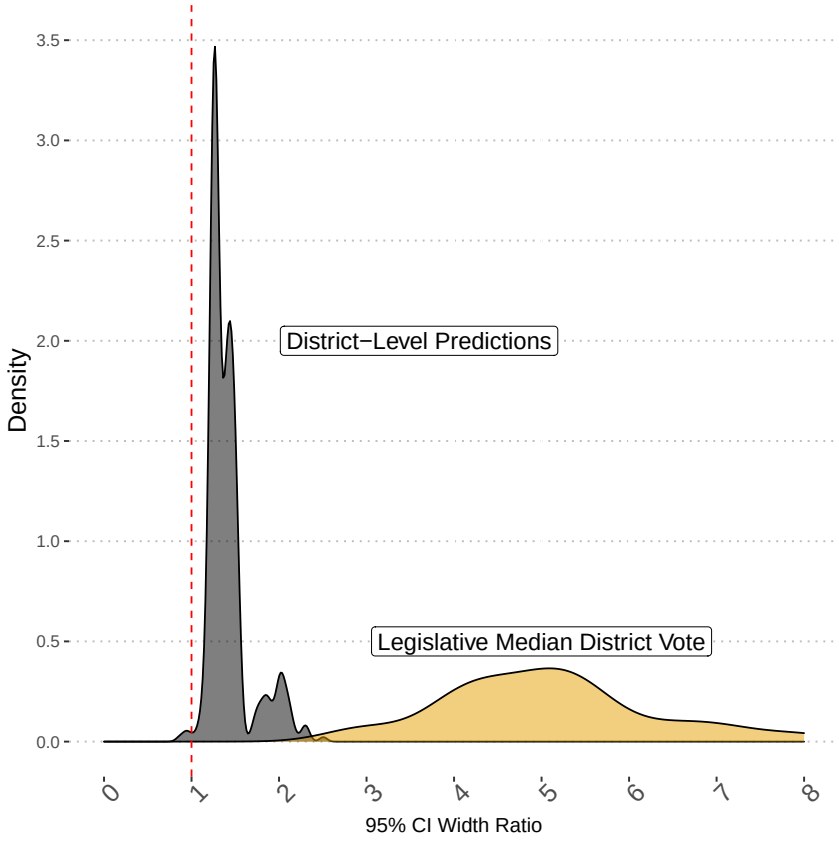


Figure 1: LogisTiCC-to-Normal Ratios of 95% Credibility Interval Widths

The linear-normal model treats the data as having 435 independent district-level observations for each election year. In reality, congressional elections data have high levels and sophisticated patterns of dependence among voting outcomes across districts. In Figure 2, we replicate the calibration exercise from Figure 3, which reports the model predictions and observed values for the median congressional seat in the given election year. We report results for three ablated models. We give the normal model with none of the modeling innovations (in gold); a model with neither a National trend assumption nor coefficient stability, but with an additive logistic student-T (ALT) assumption on the error term (in yellow); and a model with normal errors, but with a national trend and coefficient stability (in green).

We would expect a well-calibrated model to contain the true value of the median seat's vote share about ~ 95 percent of the time. To that end, we see that the normal with none of our innovations fares poorly, correctly containing the true value for the median seat only

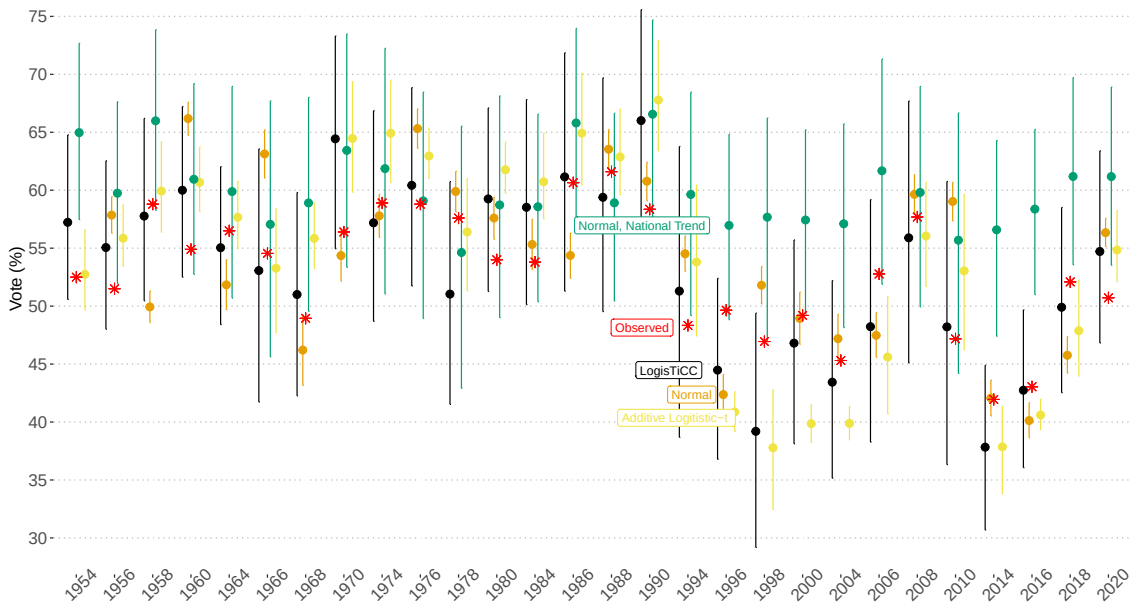


Figure 2: Comparison of Model Calibration as under Ablation

25 percent of the elections. If we switch to the ALT specification, we achieve a 40 percent accuracy rate, which is still inadequate, but better than Normal alone. When we assume normal errors with a national trend and coefficient stability, we achieve 64 percent accuracy. Under the ablated models, we find that the coefficient stability and national trend alone allow the model to achieve about 60 percent accuracy in our calibration calibration, while the ALT error assumption achieves 40 percent accuracy. Only the inclusion of all our modeling assumptions allowed us to achieve 100 percent accuracy.

In Figure 3, we reproduce Figure 3 from the paper with additional information. As in the original, the linear-normal model (in gold), which assumes independence, has confidence intervals that are extremely overconfident, and the LogisTiCC (in black) has accurately calibrated intervals. To these results, we add a version of our LogisTiCC that zeros out the parameters that model dependence. These include the national swing parameter σ_η and also our covariate stability parameter $\sigma_\beta > 0$ which, after transforming to the vote scale, also allows for some dependence across districts. In this model, we retain local uniquenesss.

Thus, we add to Figure 3, in green, estimates from the LogisTiCC model constrained to give predictions with zero cross-district independence, while retaining local unique-

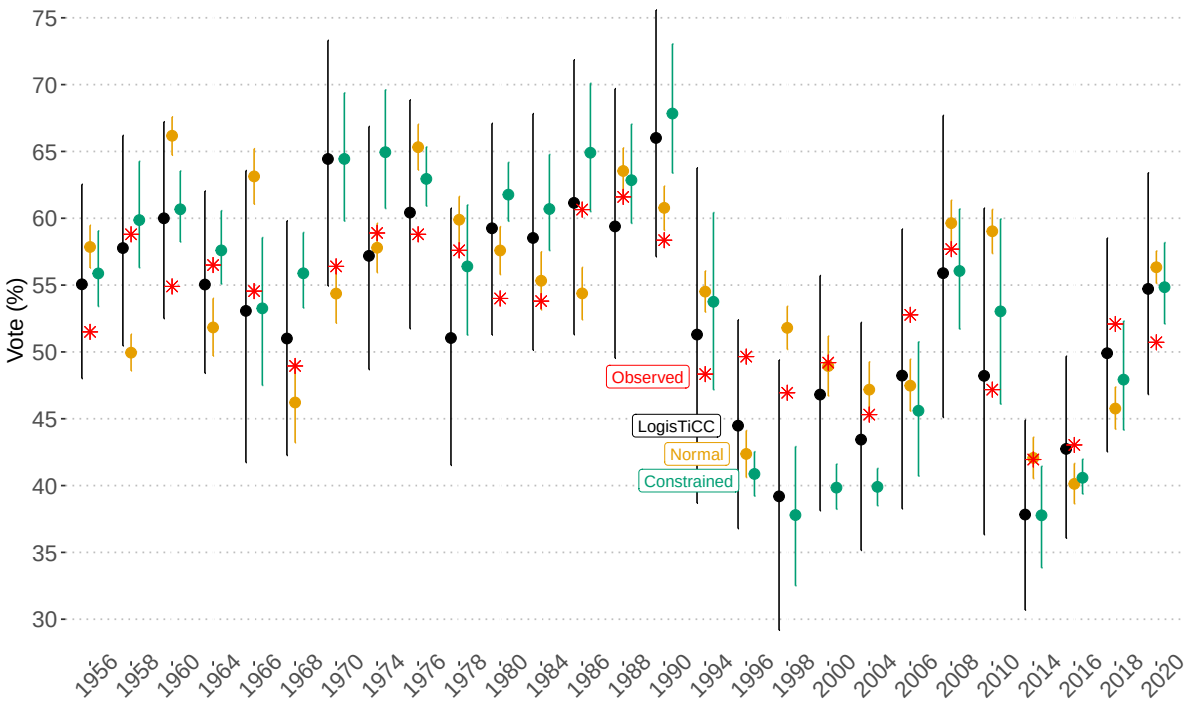


Figure 3: Expected Vote Share of the Median House Seat (95 Percent Credible Interval)

ness. While this set of assumptions reduces overconfidence of the model relative to the normal somewhat, the model is still highly overconfident. Only when we allow our full ALT error structure with cross-district correlations are the out-of-sample model predictions from the LogisTiCC well-calibrated to the historical data (in black). Under the linear-normal error structure, the incumbent party will never lose control of the House of Representative. Under the ALT without cross-district correlation, the uncertainty gets larger so that the incumbent party is sometimes forecast to lose an election, but clearly not often enough. By introducing cross-district correlation, our forecasts are well-calibrated.

3 Imputation for Uncontested Seats

Missingingess due to uncontestedness is an important feature of historical congressional election data. In Figure 4, we show the historical rate of uncontestedness in U.S. Congressional elections, which ranges from 21 percent in 1954 to 4 percent in 1996. Rather than drop these estimates which compose a nontrivial share of the data in any given election year, we impute predictive vote shares within our wholesale model framework.

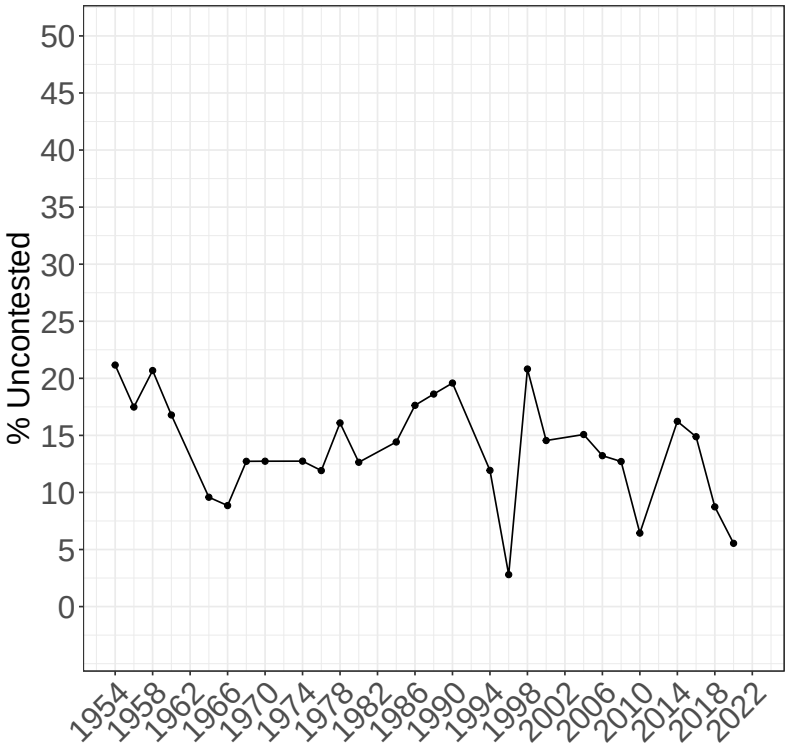


Figure 4: Uncontested Elections over Time

To account for missing data due to uncontestedness, we jointly estimate a multivariate model which predicts the uncontested vote share and missing lagged uncontested vote share. To this end, we assume that missing vote share is a censored variable where an uncontested incumbent is constrained to always win. That is, we know uncontested vote share data are not missing at random.

In Figure 5, we show that our predictions are bimodal around modes centered at 25 and 75 percent vote shares. These predictions are in line for historical estimates of uncontested vote shares.

4 Computational Details

The standard approach is usually estimated with a linear regression for forecasting (i.e., dropping γ_i) or, for other quantities of interest, via an approximate two-step procedure designed to avoid computational challenges that were difficult in the 1990s (see Gelman and King, 1994).

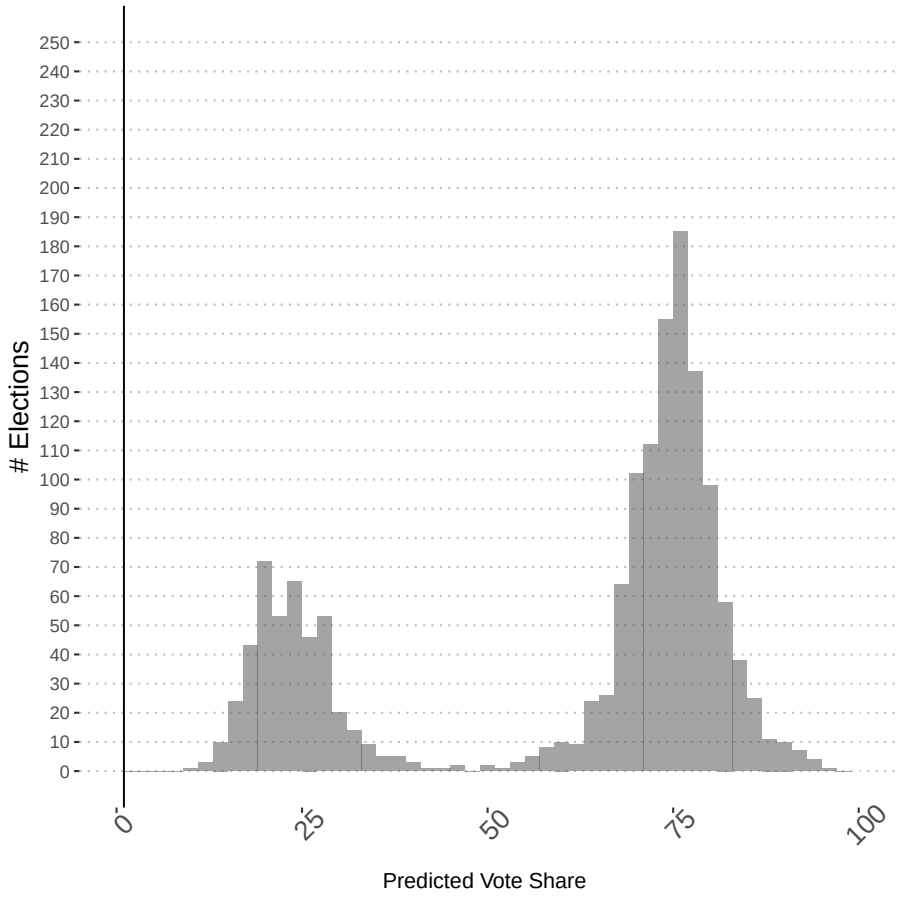


Figure 5: Histogram of Predicted Values for Uncontested Elections

Because of improvements in computation and Bayesian modeling, we estimate our LogisTiCC model via a fully Bayesian specification of Equation 2, beginning with the likelihood in Equation 5. We implement the model in "brms," open-source software that uses Hamiltonian Markov Chains (HMC) sampling to draw from the posterior distribution of a mixed-effects model (Bürkner, 2018). In practice, we draw 50,000 samples of the posterior distribution from the Bayesian mixed-effects representation. When lagged congressional vote share is a covariate, we drop the first election of each redistricting decade to fit the model. Our Bayesian methods are computationally demanding but efficient, which enables us to analyze large legislatures, and does not require asymptotic assumptions, which is especially important for legislatures like the small U.S. Senate class up for election in any one year, small national legislatures, or the many small state houses. We are also able to simulate quantities of interest directly from the full joint posterior distribution of the predicted values and parameters, which means researchers can easily

calculate any relevant quantity of interest, along with accurate and calibrated uncertainty estimates.

In order to achieve valid calibrated uncertainty estimates, we use conservative search parameters for Stan's HMC sampler. We set a delta step of 0.99, set a maximum tree depth of 10, draw $50,000$ samples with a warm up of $5,000$ iterations on 5 chains run in parallel. All Markov Chains successfully converged, with no divergent transitions, Rhats of 1 across all parameters, well-mixed chains, and no breaches of maximum tree depth.

We employ weakly informative priors for estimation convenience. In our case, because we have an average of about $1,500$ elections per decade, we do not require regularization to identify model parameters, although our weakly informative priors reduce computational time for HMC convergence. Priors are useful for speeding computation but, in our data, the choice of hyperprior parameter values does not have much effect on empirical results. The specific values we use are $\sigma_\beta, \sigma_\omega, \sigma_{tk}, \sigma_i, \sim$ Exponential(0.2) and $\nu \sim \Gamma(3, 0.5)$.

In Figure 6, we show the prior and posterior histograms for the coefficient on our predictor of the "normal" vote. This figure shows that our weakly informative prior is diffuse, while the coefficient posterior is tightly estimated around its mean, confirming that our model estimates are mostly a function of the data rather than priors. We have also found that small changes in the priors have little substantive consequences for our estimates.

Statistical results are likely less robust to the choice of the these parameters in smaller legislators. In applications with small legislatures, researchers should carefully consider the impacts of both prior specification and sampler behavior to guarantee statistically valid inference of the HMC chains.

5 Alternative Modeling Assumptions

We tried to eliminate any feature of our model not required for accurate out-of-sample validation and accurate uncertainty intervals, to include additional features that would improve performance, and to consider alternative specifications that might be easier to

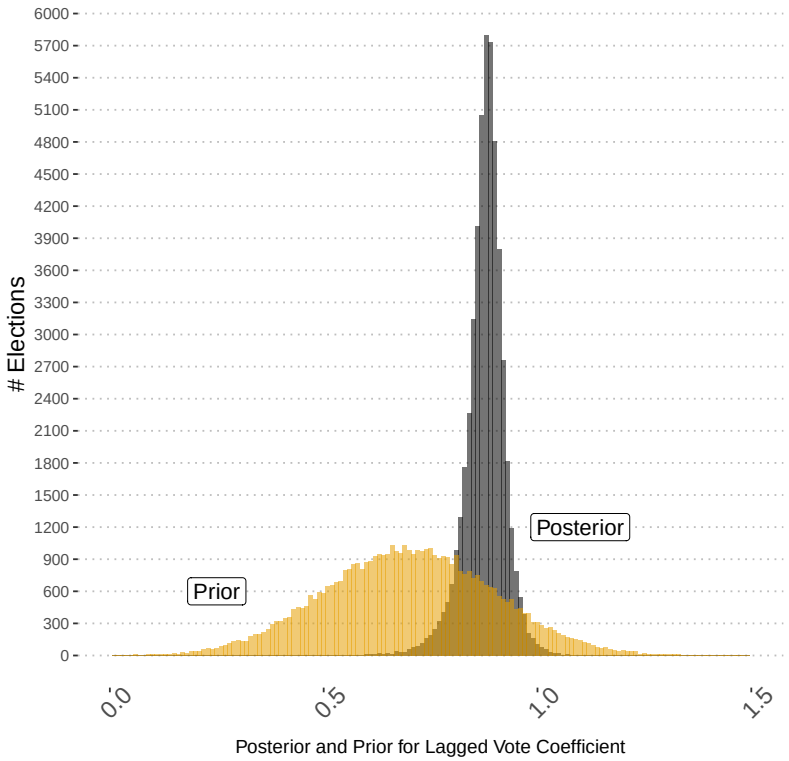


Figure 6: Posterior vs. Prior Densities

understand.

As we have shown in the main text, the linear-normal model is poorly calibrated for congressional elections. Additionally, we fit a linear-normal Student-t, which failed because it lacked the flexibility and asymmetry in the tails provided by the additive logistic t (ALT). The Additive Logistic Normal failed because it could not properly capture the levels of concentration (nearly 60 percent in the 1980s) exhibited in Figure 5a, nor did it accurately capture surprises with appropriate tails. Fitting an IID ALT, that is without contemporaneous correlations, is not well-calibrated because it misses the correlations due to year-to-year swings in the national trend or dependence due to the stability of coefficient estimates, as we showed in Figure 3.

We also tried other flexible distributions. We tried the Beta distribution, which models the unit interval directly, but produces poorly calibrated results because it, like the IID normal, does not capture appropriate levels of concentration or tail behavior. We also tried mixture distributions and errors which, while flexible, wound up being highly model

dependent, poorly identified, and computationally fragile.

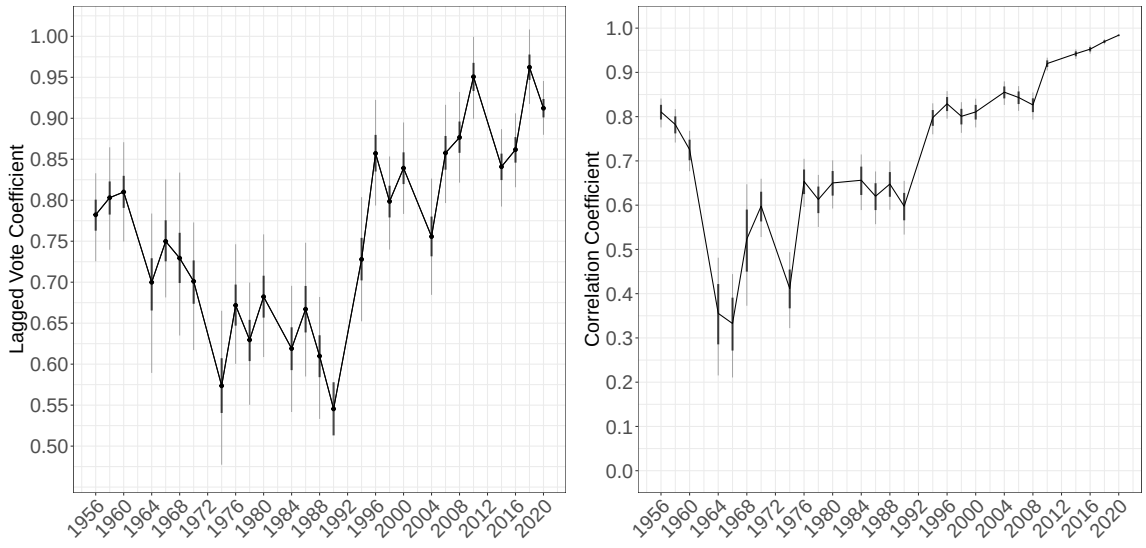
We also attempted to find alternative correlation structures, besides time mixed effects and district random effects on the logit scale, such as regional mixed effects. Besides districts in the south and outside the south, there was little predictable inter-regional variation. Districts in the North, West, and Southwest do not seem to systematically vary, conditional on the covariates. Our covariates includes an indicator for districts in the South that varies over time to capture what appear to be the most important systematic effects. In terms of covariate selection, we made choices for easy comparison to the literature. Our general model structure, like the normal, can easily accommodate other indicators when desired.

6 Additional Information about The Three Regimes

We now give additional ways of distinguishing the three regimes described in Section 4.1. These regimes are also characterized by high levels of continuity, which we convey by a plot of the coefficient on the lagged vote from our model in Figure 7a ranging in 0.8–0.95 in the early and later periods, and as low as 0.3 in the middle period. We can also see high levels of partisan alignment during the same periods outside of our model by observing the correlation between the congressional and presidential vote. We construct a time series plot of these correlations in Figure 7b, and they again reveal a now familiar U-shaped pattern.

7 The History of Generative Modeling

To calculate generatively accurate descriptive summaries, the statistical model generating these summaries should (a) pass extensive, rigorous out-of-sample tests that validate its generative abilities and (b) reflect available prior information from the literature. In our efforts to meet these conditions, we benefit from developments in three major fields of statistics, each of which has engaged with these same conditions. We now situate the ideas described in this paper (particularly Section 5) in the history of statistical analyses by briefly describing these three research traditions.



(a) Election 1–Election 2 (Lagged Vote Coefficient)

(b) Congress–President Vote (correlation)

Figure 7: Partisan Voter Alignment

First, direct attempts to build generative models in the social sciences have a long history, from path analysis originating in 1920s sociology, to linear structural equation modeling in econometrics and psychometrics in the 1960s and 70s, and, more recently, to hierarchical Bayesian models in statistics. At one point, econometricians had built many structural equation models of the economy, sometimes with hundreds of equations and each finely tuned to their past theoretical knowledge. However, rigorous out-of-sample forecasting were surprisingly embarrassed by a comparison with "atheoretical" univariate ARIMA models, leading many to reassess the value of their prior information. These attempts failed because researchers lacked the requisite computational resources to build models that reflected prior knowledge and sufficient data to make extensive validation possible. Now, model checking has become a more common part of Bayesian best practices (e.g., Gelman, Meng, and Stern, 1996).

Second, when estimating accurate generative models was not feasible, or required too many unjustified assumptions, social scientists turned to other research frameworks, often changing their quantities of interest in the process. Most notably, the literature on causal inference, especially since the 1980s, has made tremendous progress by developing ways of estimating causal effects without modeling assumptions. Although numerous articles

had previously attempted to make causal inferences, Leamer (1983) and others pointed out that high levels of (what came to be known as) model dependence meant that most of these inferences were not right, and maybe not even wrong, but instead mostly reflected researchers' priors. The "credibility crisis" that resulted from this skepticism and from rigorous tests of observational estimates compared with out-of-sample randomized experiments (Lalonde, 1986), lit a fire under the methodological community, resulting in remarkable progress that continues until today (Imbens, 2022). The theories and descriptive stories that emerge from generative models, including ours, often include many causal effects, and so the ability of these methods to proceed without modeling assumptions has been valuable for everyone. At the same time, even if we had exact knowledge of all causal effects ever estimated and a vast number of others, we would not come close to the range of descriptive knowledge social scientists seek and which can be gained by generatively accurate descriptive summaries. Descriptive quantities such as partisan bias, responsiveness, forecasts, farcasts, and many others are not causal effects but of course remain of interest to political scientists and policymakers.

Finally, machine learning methods of classification and prediction have made continual progress by their single-minded focus on out-of-sample validation. By taking their task as engineering better algorithms and downplaying constraints suggested by prior theoretical "knowledge," they make themselves continually vulnerable to being proven wrong. Although one can often do as well with simpler models that explicitly code more prior knowledge, this literature's focus on validation helps them avoid being fooled by elegant theories that do not have empirical support.

As has been true throughout the history of quantitative social science methodology, political scientists have a comparative advantage when they employ their knowledge of the political world, but do best when subjecting their statistical claims to the possibility of being proven wrong.

References

Bürkner, Paul-Christian (2018). "Advanced Bayesian Multilevel Modeling with the R Package brms". In: *The R Journal* 10.1, pp. 395–411. DOI: 10.32614/RJ-2018-017. URL: https://doi.org/10.32614/RJ-2018-017.

Gelman, Andrew and Gary King (May 1994). "A Unified Method of Evaluating Electoral Systems and Redistricting Plans". In: *American Journal of Political Science* 38.2, pp. 514–554. URL: j.mp/unifiedEc.

Gelman, Andrew, Xiao-Li Meng, and Hal Stern (1996). "Posterior predictive assessment of model fitness via realized discrepancies". In: *Statistica sinica*, pp. 733–760.

Imbens, Guido W (2022). "Causality in Econometrics: Choice vs Chance". In: *Econometrica* 90.6, pp. 2541–2566.

Lalonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs". In: *American Economic Review* 76, pp. 604–620.

Leamer, Edward E (1983). "Let's take the con out of econometrics". In: *American Economic Review* 73.1, pp. 31–43.