# Leveraging Time-Series Information to Improve Small-Area Estimation

Isaac D. Mehlhaff*

July 8, 2024

**Abstract**

Multilevel regression with poststratification (MRP) is by now the method of choice for estimating subnational quantities of interest. Scholars have proposed several methods of employing MRP on time series data, but there is limited understanding of which model specifications are most reliable under different conditions. I evaluate the effectiveness of six dynamic MRP models in estimating subnational public opinion over time. Evidence from twenty-nine policy issues reveals substantial variation in model performance, indicating the importance of careful model selection. Monte Carlo simulations using synthetic survey data show the accuracy and efficiency of each model can vary with temporal volatility in state-level opinion, the number of observed time periods, and sample size. When an appropriate model is used, dynamic MRP offers a promising method both for estimating subnational time series and for improving estimates in a single time period. I conclude with recommendations for model selection.

*The University of Chicago; imehlhaff@uchicago.edu; word count: 9,484.

Many phenomena of interest to political scientists, such as policy responsiveness or presidential voting in the United States, occur at the subnational level. However, the public opinion surveys scholars draw upon to understand them are most often designed to be representative of the national population, with too few observations to recover reliable estimates in smaller geographic units (Brace et al., 2002). Multilevel regression with poststratification (MRP) has quickly become the gold-standard method to handle this type of small-area estimation (Park et al., 2004), and many scholars have proposed further modifications to MRP to improve its performance (e.g. Bisbee, 2019; Goplerud, 2024).

At the same time, the proliferation of survey programs has endowed social scientists with an extensive library of public opinion data. These programs frequently ask the same or similar survey items in multiple years, but scholars rarely use this temporal structure to its fullest potential. In many cases, they adopt a no-pooling approach to estimate separate parameters for each year (Enns & Koch, 2013; Lewis & Jacobsmeier, 2017). This approach produces a time series, but at the expense of making strong assumptions about the evolution of subnational opinion over time. In other cases, researchers propose methods to more effectively leverage time-series information and MRP's partial-pooling benefits (e.g. Caughey & Warshaw, 2015; Franko, 2017; Pacheco, 2011). Many such methods have been proposed, but scholars currently lack a systematic understanding of which model specifications perform best, for what purposes, and under what circumstances.

I describe six possible approaches to constructing dynamic MRP models and test their ability to recover accurate subnational opinion estimates over time. Analyses of twenty-nine policy issues reveal high variation in model performance. All models produce accurate estimates on some issues and inaccurate estimates on others. The often-wide chasm between the best- and worst-performing models on a given issue should invite caution from applied researchers; without additional information about a time series' characteristics, it is difficult to prescribe *ex ante* a model to accurately estimate that time series at a subnational level.

I use Monte Carlo simulations with synthetic survey data to probe three potential sources of variation in model performance: the degree to which state-level opinion varies over time, the

number of years in which each policy preference is observed, and sample size in the target time period. I show that a no-pooling model is a poor option in many scenarios, while a model that uses random intercepts to allow the effect of individual-level demographic characteristics to vary over time offers a versatile solution.

More broadly, these analyses suggest that despite broad differences in performance observed in applied settings, dynamic MRP holds promise as a statistical method. If an appropriate model is selected for a given use case, it can provide scholars with reliable estimates of subnational opinion estimates, even in years or units where data is scarce. Even if researchers do not require time series data, they can still improve the accuracy and efficiency of their estimates in a single time period by borrowing information from other time periods, provided that opinion does not exhibit wild swings from one period to the next. As scholars continue to stress the necessity of incorporating measurement error into downstream analyses (Knox et al., 2022; Tai et al., forthcoming), these gains in efficiency may hold meaningful consequences for substantive inferences.

# Estimating Subnational Opinion from National Data Sources

Subnational politics have long been critical for understanding key concepts in political science (Giraudy et al., 2019). Key (1949) showed that states in the American South exhibited important variation in their degree of political conservatism. Arceneaux (2001) argued that public attitudes about gender roles were associated with female representation in state legislatures. More recently, scholars have drilled to the subnational level to understand democratic backsliding (Giraudy, 2015; Grumbach, 2022). Between 2004-2016, 28 percent of articles published in the *American Political Science Review* used subnational entities as the unit of analysis (Sellers, 2019).

For many years, the most common method of recovering subnational opinion estimates from national surveys was to simply disaggregate the sample into subnational units and calculate opinion within each subgroup (Brace et al., 2002; Erikson et al., 1993). However, most national surveys are too small to provide a sufficient number of observations for unbiased—let alone efficient—

estimates in subnational units. To avoid the pitfalls of survey disaggregation, scholars now frequently employ multilevel regression with poststratification (MRP; Buttice & Highton, 2013; Gelman & Little, 1997; Park et al., 2004).

MRP models of subnational public opinion use a two-stage procedure. In the first stage, survey responses are modeled hierarchically as a function of demographic and state-level variables. Adopting notation from Gelman and Hill (2007), this first-stage model often takes a form like:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i), \\ \pi_i &= \text{logit}^{-1}(\beta_0 + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{race}}_{g[i]} + \alpha^{\text{age}}_{g[i]} + \alpha^{\text{educ}}_{g[i]} + \alpha^{\text{state}}_{g[i]}), \end{aligned} \tag{1}$$

where $y_i$ is a binary survey response by respondent $i$ and $\pi_i$ gives $\Pr(y_i = 1)$ as a function of random intercepts for the respondent's gender, race, age, education, and state of residence.[1] Subscripts $g[i]$ select the random intercept for group $g$ to which respondent $i$ belongs. State intercepts are subsequently modeled as a function of state-level variables. These spatial smoothing variables are important for controlling the degree of partial pooling in the first-stage estimates and correcting for lack of representativity in cluster-sampled survey data (Buttice & Highton, 2013; Butz & Kehrberg, 2016). Here, I use the Republican share of the two-party vote in the most recent presidential election:[2]

$$\alpha^{\text{state}}_g \sim \text{N}(\gamma \cdot \text{pres}_{g[i]},\ \sigma^2_{\text{state}}). \tag{2}$$

Following Gao et al. (2021), each of the parameters in (1) and (2) take weakly informative priors:

[1] Scholars frequently refer to these parameters as "random effects." To avoid confusion with other, similarly named models, I use the "random intercepts" language recommended by Gelman and Hill (2007).

[2] Scholars often include additional state-level variables such as the percent of residents who are Evangelical Christians (e.g. Lewis & Jacobsmeier, 2017), and they frequently nest states within regions. These model features are especially useful for providing performance gains over disaggregation. Since I am primarily concerned with performance among MRP models, not between MRP and disaggregation, I opt for the comparatively simpler model.

$$
\begin{aligned}
&\beta_0,\ \gamma \sim \mathrm{N}(0,\ 2), \\
&\alpha_g^j \sim \mathrm{N}(0,\ \sigma_j^2)\ \ \forall\ \ g,\ j \in \{\text{gender, race, age, educ}\}, \\
&\sigma_j^2 \sim \mathrm{N}^+(0,\ 1)\ \ \forall\ \ j \in \{\text{gender, race, age, educ, state}\},
\end{aligned}
\tag{3}
$$

where the variances $\sigma_j^2$ are constant across groups and estimated from the data.

In the second stage, predictions from the first-stage model are calculated for each combination of demographic predictors. A final estimate of state-level opinion is then produced by calculating a weighted average of these first-stage estimates, with the joint distribution of demographic predictors in each state used as weights.[3]

Additional improvements to the MRP framework have further enhanced its performance. Leemann and Wasserfallen (2017) provide a method to relax the requirement for joint distributions of demographic predictors in the poststratification stage, Goplerud (2024) provides a fast algorithm to estimate deep interactions in the first-stage model, and several authors test the ability of machine learning techniques to improve first-stage predictions (Bisbee, 2019; Broniecki et al., 2022; Ornstein, 2020).

## Subnational Opinion over Time

The rapid advancement in MRP techniques has largely taken place by emphasizing cross-sectional estimation or, at the very least, by not allowing models to borrow information across time. With important exceptions (e.g. Caughey & Warshaw, 2015; Kastellec, 2018), the prevailing objective has been to recover a single point estimate for a subnational unit. The limited efforts to incorporate time into subnational opinion estimation frameworks is especially curious given that social scientists now have extensive access to nationally representative survey data across many years. Survey

[3]For more details on poststratification, see Leemann and Wasserfallen (2017), Park et al. (2004), and Warshaw and Rodden (2012), among others.

programs often capture opinion on the same topics over a long period of time, including attitudes on specific policy domains (Caughey & Warshaw, 2018) or more general measures of ideology or partisanship (Mehlhaff, forthcoming).

Some MRP researchers have already noted this opportunity. Lax and Phillips (2009a) measure state-level support for gay rights by combining 41 polls fielded from 1999-2008. Their model includes a random intercept for each poll, but they do not break down the results by year. On one hand, this complete-pooling approach has the benefit of dramatically increasing the number of observations available to the model,[4] perhaps leading to more accurate estimates in small states and more efficient estimates overall. On the other hand, it implicitly assumes that state-level opinion is stable over the time period under consideration and, more specifically, that the *relationships* between demographic predictors and attitudes are stable over time. In many cases, at least one of these assumptions is unlikely to hold.

Ideally, MRP models of dynamic processes would incorporate time more explicitly. Doing so could impart at least two benefits. First, preserving the temporal dimension in opinion can be critical to answering causal questions (Blackwell, 2013). When research topics do not lend themselves to experimentation, leveraging temporal variation is often one of the few ways researchers can draw causal inferences from aggregate-level observational data. Even where causal inference is not appropriate, temporal variation can enhance external validity or help assess theoretical scope conditions. Second, incorporating data from additional time periods can improve cross-sectional MRP estimates in years or units where data is scarce. By borrowing information from other time periods, MRP's partial pooling benefits may enable the model to more accurately and efficiently estimate cross-sectional opinion.

Two recent examples demonstrate the research questions that can be asked and answered with dynamic MRP. Smith et al. (2020) create time-varying, state-level estimates of racial resentment. Using descriptive analyses of how racial animus varies across states and how it changes within

[4]My focus in this paper is on using MRP for time series estimation. I therefore use "complete-," "partial-," and "no-pooling" terminology to refer to how the model handles *temporal* information. All models use a partial pooling structure for spatial and demographic variation.

states over time, they challenge the narrative of racial progress and show that many states—even outside the South—have increased in racial resentment since the 1980s. Claassen and Traunmüller (2020) show how dynamic MRP can help scholars estimate population-level quantities like religious demographics that are rarely included in census questionnaires. They measure the population of Muslims, Hindus, and Jews in the United Kingdom over two decades, both overall and in detailed demographic subgroups. These two examples are also illustrative because the two authorship teams use quite different approaches to incorporating time into their MRP models; Smith et al. (2020) use random intercepts by year, while Claassen and Traunmüller (2020) use several variations of what I will refer to below as a "linear trend" model. Which method is more appropriate? Do different methods lead to different results?

When these and other authors implement dynamic MRP methods, they typically attempt to validate their estimates in their specific use cases. However, these validity checks are largely evaluations of covariate selection or interaction effects within one general model type. At present, the literature provides little guidance on how well the models themselves perform, which ones are appropriate for which objectives, and under what conditions dynamic MRP can be reliably employed.

One notable exception is a manuscript by Gelman et al. (2018). They propose eighteen possible models, varying the type of covariates included as well as the manner in which time is introduced into the model. Testing these models on survey data capturing the support for same-sex marriage from 1993-2004, they argue that dynamic MRP is particularly helpful for increasing the accuracy of estimates in units where data is scarce. They specifically recommend using time as a continuous smoothing variable and discourage year or demographic-year random intercepts. However, by analyzing only one time series, they leave open questions as to how performance varies with factors such as the degree of over-time variability in opinion or the number of time periods under consideration. While I concur with Gelman et al.'s emphasis on sparse data situations, I push back on their modeling recommendations. Evidence from survey and synthetic data below shows that mod-

els with a time-smoothing parameter perform quite poorly, while models that allow demographic random intercepts to vary over time frequently outperform other approaches.

## Approaches to Dynamic MRP

There are many ostensibly reasonable model specifications for dynamic MRP, including combinations of the candidate models I examine here. For simplicity, I focus on six distinct methods of estimating subnational opinion over time; scholars using dynamic MRP in applied settings should systematically evaluate whether adding more complexity on top of these foundational models improves performance in their specific use case.

The simplest approach is a no-pooling model that estimates completely separate models by year and allows no data sharing among them. This model is mathematically identical to the complete-pooling model in equations (1) through (3), but it is fit once for each year, resulting in $T$ total models, each producing estimates for year $t \in \{1, ..., T\}$. This is a popular approach in applied work using MRP to produce time series estimates (Butz & Kehrberg, 2016; Enns & Koch, 2013; Lewis & Jacobsmeier, 2017). Its output no doubt adheres more closely to ground-truth opinion trends, but it also prohibits the model from using any information from other years to estimate opinion in the target year—all but eliminating the partial-pooling benefits of MRP and risking imprecise estimates (Caughey & Warshaw, 2019)—and it makes arbitrary assumptions about the degree to which opinion changes over time (Gelman et al., 2018). Particularly in states or years with small sample sizes, there is likely room to improve performance by allowing the model to borrow information across time periods.

The other five candidate models generally seek to leverage the partial-pooling benefits of MRP, incorporating richer data than a no-pooling model can but shrinking yearly estimates to the appropriate means when data is scarce or variation across years is negligible. Pacheco (2011, 2014) offers one simple, albeit rather inflexible, method of achieving partial-pooling across time. She pools data within three- and five-year intervals and estimates the model above on each moving

window of data. In effect, this procedure is a no-pooling model with some manually imposed amount of partial-pooling, determined not by the model but by the researcher. It may outperform a pure no-pooling model by borrowing information across a small set of contiguous years, but pooling outside each moving window is impossible, again ensuring that the model does not see most of the available data. Moreover, this approach assumes—similar to complete-pooling—that opinion does not change within each moving window.

A more adaptable strategy would be to incorporate time into the model itself. The model in (1) can be adjusted to include a continuous time-smoothing variable:

$$\begin{aligned} \pi_i &= \text{logit}^{-1}(\beta_0 + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{race}}_{g[i]} + \alpha^{\text{age}}_{g[i]} + \alpha^{\text{educ}}_{g[i]} + \alpha^{\text{state}}_{g[i]} + \delta \cdot \text{year}), \\ \delta &\sim \text{N}(0,\, 2). \end{aligned} \tag{4}$$

All other terms are defined as in (2) and (3). Time-smoothing is a popular strategy in applied work (Shirley & Gelman, 2015; Wiertz & Lim, 2021), but it has the disadvantage of assuming linear time trends. Some authors have allowed for more flexibility by including a second-order polynomial (Kastellec, 2018), by estimating separate slopes for each demographic category (Claassen & Traunmüller, 2020), or by using splines to determine the functional form for time (Kołczynska et al., 2024). The latter solution may especially hold promise in cases where opinion changes rapidly from year to year.

A much simpler alternative to fitting splines would be to follow the same structure as the classic MRP model in (1) and use random intercepts for each year, just as the model uses random intercepts for demographic groups and states. This is also a common method of estimating dynamic MRP models in applied research (Ben-Shalom et al., 2021; Simonovits & Bor, 2023; Smith et al., 2020).[5] Adding random intercepts by year to the baseline model in (1) requires specifying two additional priors:

[5]Merely including these random intercepts does not, by itself, make the output dynamic. Lax and Phillips (2009b) and Warshaw and Rodden (2012) include random intercepts by year in their models but do not poststratify by year; although the year of the survey is used to improve first-stage model fit, they do not use it to produce dynamic estimates.

$$
\begin{aligned}
\pi_i &= \text{logit}^{-1}(\beta_0 + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{race}}_{g[i]} + \alpha^{\text{age}}_{g[i]} + \alpha^{\text{educ}}_{g[i]} + \alpha^{\text{state}}_{g[i]} + \alpha^{\text{year}}_{t[i]}), \\
\alpha^{\text{year}}_t &\sim \text{N}(0,\ \sigma^2_{\text{year}})\ \ \forall\ t, \\
\sigma^2_{\text{year}} &\sim \text{N}^+(0,\ 1).
\end{aligned}
\tag{5}
$$

The key benefit of partial-pooling in this case comes from allowing the model to estimate $\sigma^2_{\text{year}}$ from the data. A no-pooling model effectively assumes $\sigma^2_{\text{year}} \to \infty$, which is unlikely to be reasonable in most dynamic applications.

One important limitation of a model with random intercepts by year is that the time component is not allowed to interact with any other terms, requiring the assumption that the effect of each demographic category on the dependent variable is constant over time (Ben-Shalom et al., 2021). The model in (5) would therefore be hard-pressed to capture features like, for instance, an expanding gender gap in political attitudes (Clark, 2017). This assumption can be relaxed by not using separate random intercepts by year but rather by allowing the demographic effects to vary by year in the form of demographic-year random intercepts:

$$
\begin{aligned}
\pi_i &= \text{logit}^{-1}(\beta_0 + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{race}}_{g[i]} + \alpha^{\text{age}}_{g[i]} + \alpha^{\text{educ}}_{g[i]} + \alpha^{\text{state}}_{g[i]} + \alpha^{\text{year}}_{t[i]} \\
&\qquad + \alpha^{\text{gender}}_{g[i],t[i]} + \alpha^{\text{race}}_{g[i],t[i]} + \alpha^{\text{age}}_{g[i],t[i]} + \alpha^{\text{educ}}_{g[i],t[i]} + \alpha^{\text{state}}_{g[i],t[i]}), \\
\alpha^{\text{state}}_{g,t} &\sim \text{N}(\gamma \cdot \text{pres}_{g[i],t[i]},\ \sigma^2_{\text{state}}), \\
\alpha^{j}_{g,t} &\sim \text{N}(0,\ \sigma^2_j)\ \ \forall\ g,\ t,\ j \in \{\text{gender, race, age, educ}\},
\end{aligned}
\tag{6}
$$

where subscripts $g[i],t[i]$ select the demographic group and year, respectively, to which each observation belongs. I retain the random intercepts from (5) to enable partial pooling for both demographic and year estimates. $\beta_0$ and $\gamma$ are defined as in (3) and $\sigma^2_j \ \ \forall \ \ j \in \{\text{gender, race, age, educ, state}\}$ are constant across groups and years and estimated from the data.

This approach mirrors the practice of allowing state random intercepts to vary by year (Ben-Shalom et al., 2021; Shirley & Gelman, 2015), but extends it to all demographic variables.

Fitting separate random intercepts for each demographic-year combination may help the model capture large swings in opinion over time. But in some ways, it still relies on the same unrealistic assumption that a no-pooling model does: that time is merely a collection of years, where opinion can be measured independently in each year. In reality, opinion at time $t$ depends on opinion at time $t-1$ but not on opinion at time $t+1$. Therefore, it might be beneficial to impose a greater degree of structure on the over-time changes in demographic-year intercepts.

Gao et al. (2021) show that directed structured priors on individual-level MRP parameters, such as age, lead to bias and variance reduction in first-stage estimates. I adapt their approach for dynamic applications, placing a local-level transition model on the year and demographic-year random intercepts from (6):[6]

$$
\begin{aligned}
\pi_i = \text{logit}^{-1}(&\beta_0 + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{race}}_{g[i]} + \alpha^{\text{age}}_{g[i]} + \alpha^{\text{educ}}_{g[i]} + \alpha^{\text{state}}_{g[i]} + \alpha^{\text{year}}_{t[i]} \\
&+ \alpha^{\text{gender}}_{g[i],t[i]} + \alpha^{\text{race}}_{g[i],t[i]} + \alpha^{\text{age}}_{g[i],t[i]} + \alpha^{\text{educ}}_{g[i],t[i]} + \alpha^{\text{state}}_{g[i],t[i]}), \\
\alpha^{\text{year}}_t \sim \text{N}(&\alpha^{\text{year}}_{t-1}, \sigma^2_{\text{year}}), \\
\alpha^j_{g,t} \sim \text{N}(&\alpha^j_{g,t-1},\ \sigma^2_j)\ \ \forall\ g,\ t,\ j \in \{\text{gender, race, age, educ}\}.
\end{aligned}
\tag{7}
$$

All other terms are defined as above, including $\alpha^{\text{state}}_{g,t}$, which does not take a dynamic model but rather is identical to its implementation in (6). This approach models time-dependent parameters with a random walk, making them a function of the parameter's value in the previous time period plus random noise. It has the effect of shrinking the posterior of, for example, $\alpha^{\text{year}}_t$ toward the posterior of $\alpha^{\text{year}}_{t-1}$. This structured prior enables complete information-sharing among years, under

[6] See also Caughey and Warshaw (2015, 2018). Their dynamic group-level item response model effectively reduces to a dynamic MRP model in the special case when there is only one survey item.

the assumption that individuals in each demographic cell have similar opinions to individuals with the same demographics in previous years.

Finally, local-level transition models like this one must be anchored to some baseline value. I do this by placing a prior on the first estimate in each time series:

$$\alpha^{j}_{g,1} \sim \mathrm{N}(0,\ 1)\ \ \forall\ \ g,\ j \in \{\text{gender, race, age, educ}\}. \tag{8}$$

I fit all models in a fully Bayesian framework, which enables me to produce uncertainty estimates as a direct byproduct of the model-fitting process. These are important for assessing the efficiency of each modeling strategy relative to the others. I provide more details on estimation in Supplementary Information (SI) section A.

## Dynamic MRP on Twenty-Nine Policy Issues

Testing the performance of the six models in the previous section requires data from large national surveys that consistently ask respondents the same items over multiple years. I use the Cooperative Election Study (CES), which consistently asks a large, Census-benchmarked sample of Americans for their views on policy issues. I take twenty-nine individual time series from these data, on issues ranging from abortion and healthcare to immigration and military policy.[7] The CES also provides the four demographic variables I included in the models above, which I supplement with state-level information about Republican vote share in the previous presidential election. This broad collection of public opinion data enables me to test dynamic MRP models in a diverse set of time series. If one model specification is systematically superior to the others, it should be evident in this analysis.

Importantly, the CES is a very large national survey, with a mean sample size of 34,815 respondents per year. This allows me to use disaggregated state-year averages, weighted to be representa-

[7] I limit my search to survey items included in at least four consecutive years and are either asked on a binary scale (e.g. support or oppose) or on an ordinal scale that could be converted to binary without excluding a middle category. SI section B.1 provides survey item details.

tive of the population, as the benchmark (Buttice & Highton, 2013). Lax and Phillips (2009b) point out that disaggregation and survey weights may both bias results against MRP. Because I am not interested in comparing MRP to other methods but rather comparing across different *approaches* to implementing MRP in a time series application, it primarily matters that I compare all candidate models to the same baseline—how the baseline is constructed or whether poststratification should be used in combination with survey weights matters relatively little in this case (see also Bisbee, 2019). Since survey weights tend to bring opinion estimates closer to the "ground truth," I opt to use them.[8]

I randomly sample ten percent of responses in each year to fit the MRP models and use data from the United States Census to poststratify estimates (Ruggles et al., 2024). I assess model performance by calculating the root mean squared error (RMSE) across all state-years within each policy issue.[9] Figure 1 displays these error metrics, with differently colored points representing each model. Policy issues are ordered along the y-axis according to the RMSE of the no-pooling model. The bolded y-axis label denotes the average model RMSE across all issues.

Three conclusions stand out from this analysis of twenty-nine time series. First, in line with Buttice and Highton's (2013) finding, model accuracy varies substantially across issues. Most RMSE metrics cluster between five and seven percentage points, but estimates for some issues are consistently accurate across all models while others are consistently poor. RMSE for all models on using the military to spread democracy are around 0.04; error on the same-sex marriage issue is more than double that, at 0.08 and above.

Second, no one model consistently outperforms the others. Models with random intercepts by demographic-year have the lowest error when averaging over all issues, while the local-level transition model is slightly more accurate than a no-pooling model. However, these averages belie substantial variation. Every model claims the lowest RMSE on at least one issue. At the same time, the model with demographic-year random intercepts is the only one that does *not* claim the

[8] In SI section B.2, I show that substantive conclusions differ little when using an unweighted baseline.

[9] SI section B.3 displays alternative performance metrics such as MSE, MAE, correlation, and standardized bias (Buttice & Highton, 2013).

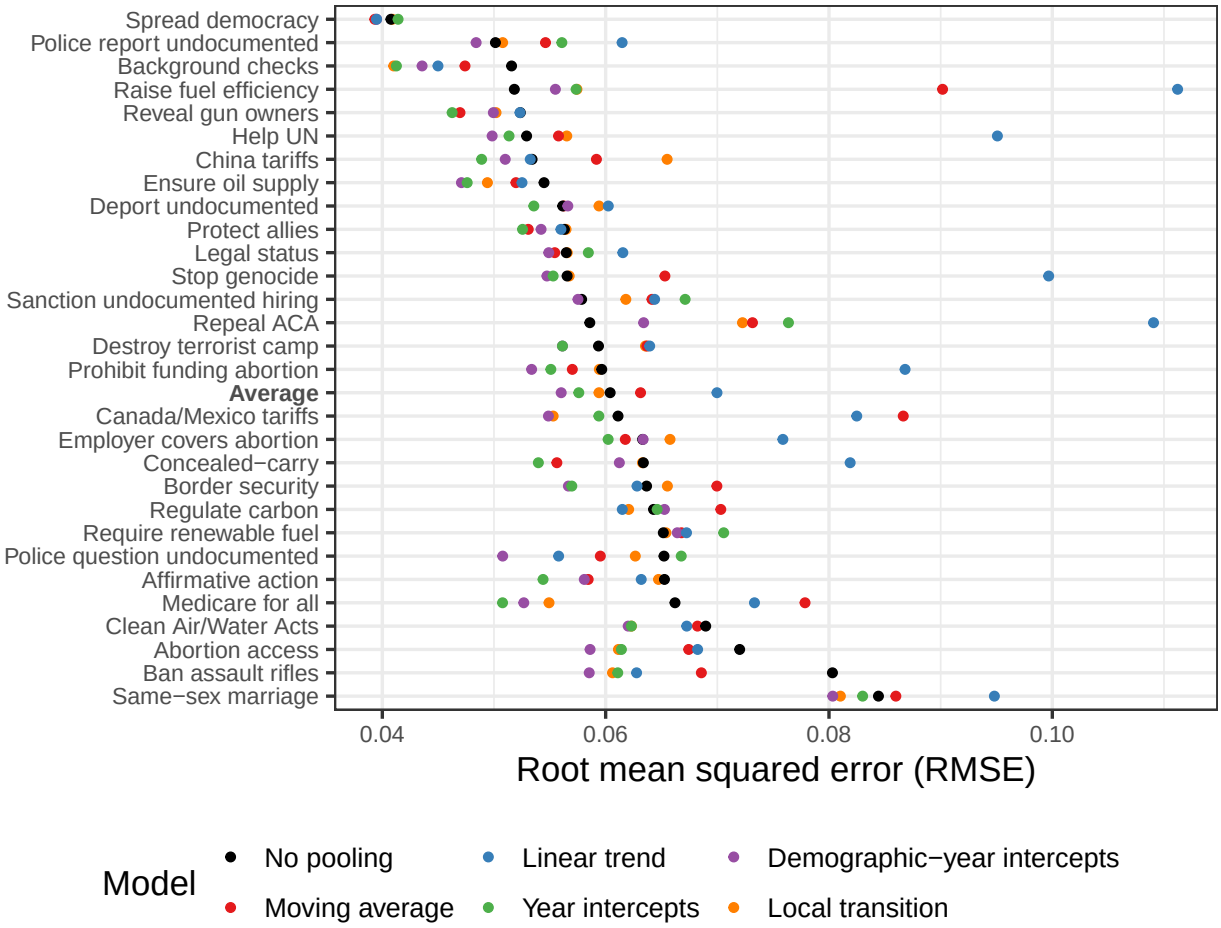


Figure 1: RMSE of CES Time Series Estimates. Performance metrics reflect model estimates for all state-years within each policy issue. Policy issues ordered on y-axis according to no-pooling RMSE.

highest RMSE on at least one issue. Without additional knowledge on the factors that cause some dynamic MRP models to perform better or worse, it is difficult to identify one model that is likely to provide high-quality estimates.

Third, and perhaps most alarming, the gap between the best- and worst-performing models is occasionally small but more often vast. Evidenced by the points scattered across the right half of Figure 1, many models return error rates at least double the best-performing model on each issue. Linear trend models are disproportionately represented among these low-accuracy outliers. The clearest example is the issue of requiring automobile manufacturers to increase the fuel efficiency of their cars. A no-pooling model exhibits RMSE of about 0.05 on this item, but the linear trend

model's RMSE is over 0.11. Choosing the wrong model for a particular application may lead to inferences that are inaccurate, inefficient, or both.

## Explaining Variation in Dynamic MRP Performance

The previous section showed wide variation in model performance across an equally wide-ranging set of time series. Scholars have identified several sources of variation in MRP performance, such as sample size (Lax & Phillips, 2009b), multi-level model complexity (Warshaw & Rodden, 2012), or the importance of individual- and state-level covariates (Buttice & Highton, 2013). As this is a dynamic application of MRP, I focus on three characteristics of the time series themselves that I expect to contribute to model performance.

First, when opinion is highly volatile over time, models that conduct more temporal smoothing—like those with linear trends or local-level transitions—may lose out to no-pooling or moving average models. Notably, the policy issue in Figure 1 with the most accurate predictions overall—using the military to spread democracy—is also the most stable over time; the standard deviation of states' opinion on this policy across years is less than four percentage points on average.

Second, the length of the time series may affect each model differently. No-pooling and moving average models produce completely separate estimates for each year, but hierarchical models like those with random intercepts rely on being able to borrow information across many time periods. More complex models may require even more data to fit their larger inventory of parameters. The "abortion access" issue, on which no-pooling performs markedly worse than the local-level transition, has sixteen years of data. Other issues have as few as four.[10]

Finally, just as static MRP tends to perform better in larger states with larger sample sizes, certain dynamic MRP models are likely sensitive to small sample sizes in some years. Although the average CES sample is close to 35,000 observations, there is meaningful variation around that

[10]SI section B.4 provides the standard deviation of over-time opinion within states and the number of available years of data.

mean. In 2020, the CES conducted 61,000 interviews. In 2007, it conducted only 9,999. Some dynamic models might perform worse when the time series includes 2007 than when it includes 2020. In the next section, I assess these eventualities by using Monte Carlo simulations that allow me to systematically vary key aspects of the data-generating process.

# Simulation Evidence

Most MRP advances are validated on data from large national surveys or state polls (Bisbee, 2019; Warshaw & Rodden, 2012). This has the benefit of testing how models perform when data contains the type of noise one would expect from real-world data-generating processes. However, Monte Carlo simulations are particularly useful for testing complex models like MRP for at least two reasons. First, the performance of MRP models depends on many different sources of variation. A simulation approach gives me complete control over the data-generating process, allowing me to hold constant sources of variation which I do not wish to test—such as the degree to which state-level variables predict opinion—and randomize the sources of variation in which I am interested, like the variation in state-level opinion over time.

Second, to establish a ground truth benchmark with real-world survey data, researchers must either disaggregate the full sample into state-level means (Lax & Phillips, 2009b) or draw on separate data sources such as state-level surveys or election returns (Park et al., 2004). Both are suboptimal. Survey disaggregation—which I used in the previous section out of necessity—requires extremely large samples and is the very method known to produce inefficient and often biased estimates. State-level surveys are rare, meaning that not every state-year may be verifiable. Election returns are not representative of the state population. By contrast, Monte Carlo simulations ensure that I know the true value of opinion in each state-year.

I use synthetic survey data to evaluate the performance of each MRP model in three different contexts. In each simulation, I calculate the RMSE as an overall error metric and I perform the

bias-variance decomposition to assess accuracy and efficiency separately.[11] This decomposition can be important in model selection. Two models may have similar RMSE values even though one produces precisely estimated, incorrect values while the other produces unbiased estimates with very high uncertainty. When a bias-correction adjustment is feasible, scholars may prefer the former. When model estimates are primarily used for descriptive purposes and not as variables in downstream analyses, they may prefer the latter.

These simulation exercises produce three main takeaways. First, efficiency and accuracy of subnational time series estimates varies across models and with the degree of over-time volatility in state opinion. The popular no-pooling approach is never the best option and, when opinion is more stable, is often the worst. Second, models with demographic-year random intercepts achieve consistently high performance as the length of the time series increases, regardless of whether state-level opinion is stable or volatile. No other model produces such consistent performance, and some simpler models can get worse when they see more time periods. Third, the demographic-year random intercepts model also offers a versatile method to recover accurate state-level estimates in years where data is scarce. Other models require larger sample sizes and are only viable options under conditions of strong temporal stability. In SI section C, I provide an analysis of computational efficiency and illustrate that although more accurate models do entail higher memory usage, they do not always require more time to fit.

## Generating Synthetic Survey Data

The simulated data consist of $S = 10$ states,[12] each observed at $T = 10$ time periods. The total population contains $N = 10,000$ survey respondents who are assigned to gender, race, age, and education categories in proportion to the prevalence of those categories in United States Census data. Respondents are divided equally among states and observed in each time period $t \in \{1, ..., T\}$.

[11] $\text{RMSE}(\hat{\theta}) = \sqrt{\text{Bias}^2(\hat{\theta}, \theta) + \text{Var}(\hat{\theta})}$, deriving from the more common $\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}, \theta) + \text{Var}(\hat{\theta})$ for an estimator $\hat{\theta}$. RMSE is measured in the same units as the predictions, so I present it instead of MSE to aid interpretation.

[12] States could also be understood as legislative districts or any other subnational geographic entity.

The next step is to assign a state-level covariate. In keeping with the model specifications above, I conceptualize this as the Republican share of the two-party vote: $\text{pres}_{s,t} \sim \text{Beta}(20,\ 20)$. The data therefore consists of 10,000 synthetic survey respondents in each time period, with demographic characteristics that match the United States in the aggregate. These respondents are nested in 10 hypothetical states, each with a randomly assigned presidential vote share that changes each time period.

The final piece of data I need to generate is the binary dependent variable $y$. Each observation $y_i$ is an independent draw from a Bernoulli distribution, where the probability of selecting the positive category, $\pi$, varies across states and time periods:

$$y_{i[s,t]} \sim \text{Bernoulli}(\pi_{s,t}). \tag{9}$$

In this context, $\pi_{s,t}$ also represents the true, population-level opinion in state $s$ at time $t$—the quantity I aim to estimate with each model. Buttice and Highton (2013) fix $\pi_{s,t} = 0.5$ in their simulations, but I seek to understand how model performance varies with the over-time volatility of state-level opinion. I therefore allow $\pi_{s,t}$ to vary by randomly drawing $\pi_{s,t}$ from a beta distribution, whose parameters are themselves randomly determined at each iteration:

$$\begin{aligned} \pi_{s,t} &\sim \text{Beta}(\psi, \omega), \\ \psi = \omega &\sim \text{U}(0.1, 100). \end{aligned} \tag{10}$$

$\psi$ and $\omega$ are constant across states and time periods. In each iteration, therefore, the simulation draws a pair of beta parameters and uses them to generate a true value of opinion in each state and time period. When $\psi$ and $\omega$ are closer to 100, the beta distribution in (10) will be more tightly centered around 0.5 and state-level opinion will be very stable over time. When $\psi$ and $\omega$ are closer to zero, the beta distribution will be more diffuse and opinion will vary dramatically from one time period to the next.

I conduct 300 iterations, each using the process above to generate synthetic population-level data. Within each iteration, I randomly sample ten percent of the generated data, fit each model on the sampled data, and calculate performance metrics, taking $\pi_{s,t}$ as the true value of opinion in each state and time period.

Some features of the simulation are clearly not representative of most real-world data sources. For example, I allocate respondents to states in equal proportions and I essentially ensure that covariates have negligible associations with the dependent variable. These decisions are driven by my desire to isolate variation in temporal opinion trends and assess their effect on model performance. Buttice and Highton (2013) construct their Monte Carlo simulations by explicitly imposing a covariance structure between covariates and the dependent variable. However, I am neither interested in how well each model can recover this covariance, nor in how variations in covariance affect downstream model performance. Therefore, instead of arbitrarily choosing values for these parameters, I opt to completely randomize them and hold them constant from one simulation to the next.

Also noteworthy is my decision to avoid imposing over-time trends and instead allow opinion to fluctuate randomly. This simulation therefore favors a no-pooling model at lower levels of $\psi$ and $\omega$, as each state-year is essentially independent of the others in the data-generating process. The no-pooling model's comparative advantage diminishes somewhat at high levels of $\psi$ and $\omega$, at which point the simulation produces a linear, flat time trend. In these cases, more structured models, such as the linear trend and local-level transition models, are more aligned with the data-generating process. It is important to keep these structural advantages in mind when evaluating model performance.

## Estimating Time Trends in State-Level Opinion

I turn first to an analysis of each model's ability to accurately recover the simulated time series as a whole. This provides a general overview of model performance in the most common use case for dynamic MRP. Figure 2 displays the bias, variance, and RMSE of each model's estimates, plotted

against the randomized values of $\psi$ and $\omega$ from (10).[13] Higher x-axis values therefore indicate more stability in opinion trends over time. Lines are plotted using locally estimated scatterplot smoothing (LOESS) and display 95 percent confidence intervals.

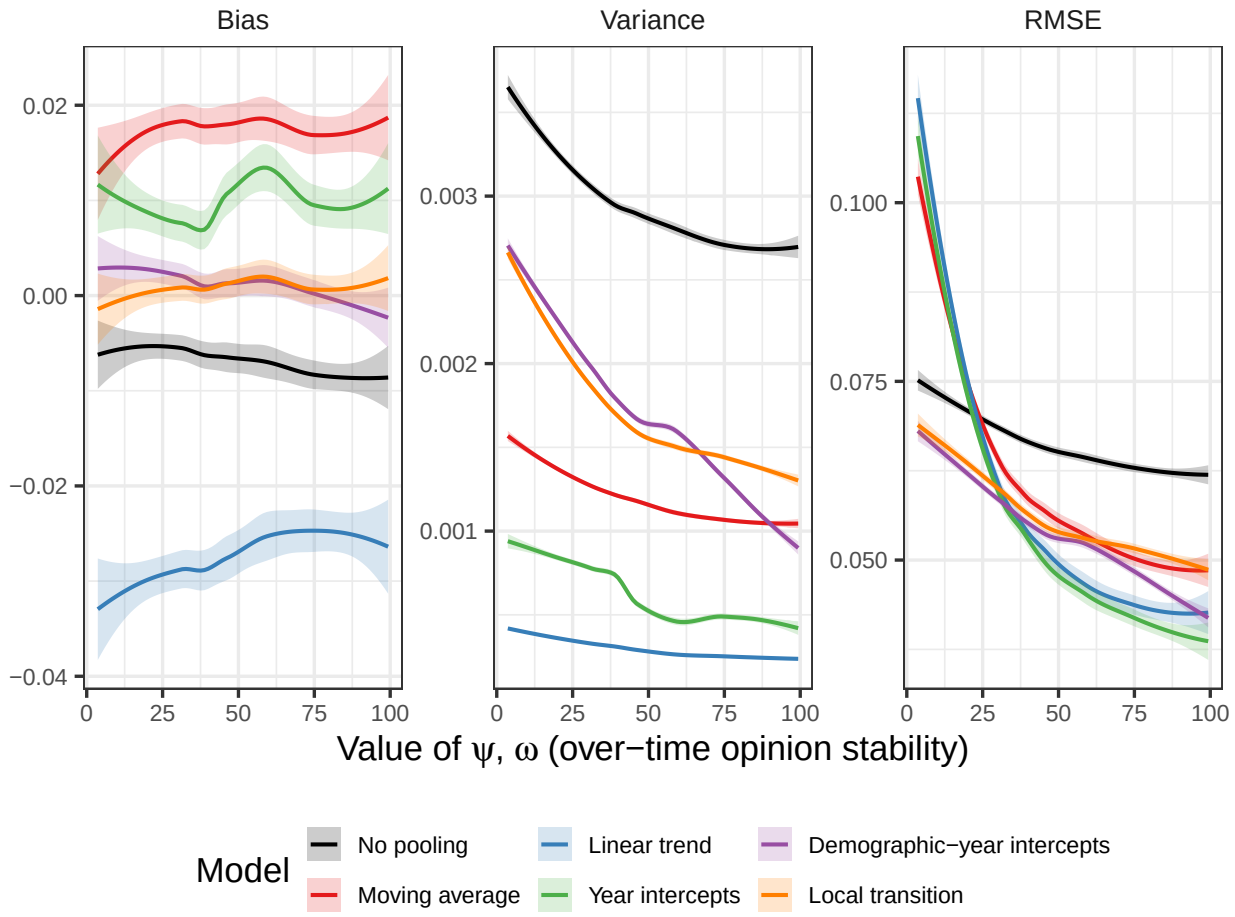


Figure 2: Bias, Variance, and RMSE of Time Series Estimates. Performance metrics reflect model estimates for all states and time periods within each iteration. Higher values of $\psi$ and $\omega$ indicate more stability in opinion trends. Trend lines show 95% confidence intervals.

Calculating metrics across all state and time-periods, the models allowing demographic effects to vary over time (demographic-year intercepts and local-level transition models) tend to perform best, returning virtually unbiased estimates even in this finite-sample situation. Somewhat sur-

[13] Bias $= \mathbb{E}(\hat{\theta}) - \theta$ for an estimator $\hat{\theta}$. A trend line close to zero could therefore conceal variation across individual states and time periods, as positive and negative biases would effectively cancel each other out when calculating the line of best fit. Large confidence intervals on the trend line can help indicate when this is occurring. I omit the squared bias, as it is very similar to RMSE.

prisingly, the no-pooling model exhibits lower bias than moving average or linear trend models at all levels of opinion volatility, though the linear trend model appears to improve slightly at lower levels of volatility. As might be expected owing to the bias-variance tradeoff, the variance of the estimates shows approximately the opposite pattern. The linear trend is by far the most error-prone but introduces very little uncertainty, while the no-pooling model contains the greatest degree of variation in its estimates. Most models get more efficient as opinion becomes more stable over time, with the demographic-year intercepts model showing an especially precipitous decline in variance.

The plot of RMSE combines the first two plots to provide a summary measure of performance. In some respects, this simulation shows that volatility in over-time opinion carries implications for model selection. Models employing moving averages, linear trends, or random intercepts by year perform poorly when opinion varies substantially over time, with RMSE values nearly double that of the best-performing model. These poor results cannot be completely explained away as artifacts of the data-generating process; a moving-average model is most similar to a no-pooling model while a linear trend model is highly structured, much like a local-level transition model. However, the no-pooling and local-level transition models exhibit much lower RMSE when $\psi, \omega < 25$. The standard deviation of a distribution generated by this value of $\psi$ and $\omega$ is about 0.071, still lower than the over-time variation in a quarter of the CES opinion trends from the previous section.

At more stable levels of opinion, the performance of most models incorporating time converge, and all outperform a no-pooling approach. This makes intuitive sense; when opinion swings wildly from one time period to the next, dynamic models are mostly fitting to noise. In this case, the no-pooling model's arbitrary assumptions about the relationship between time periods is much less consequential. But as opinion at time $t$ becomes more tightly coupled to opinion at time $t-1$, allowing the model to learn from that relationship can bring noticeable gains.

## Increasing Time Series Length

In addition to the degree of opinion volatility, model performance could also vary with the number of time periods available. To assess this possibility, I focus specifically on each state's opinion at $t_1$. I start by fitting each model with only this one time period of data and gradually add time periods one by one until the full time series is included. I only calculate performance metrics on $t_1$ because this allows me to hold the target values constant. If I added each subsequent time period as $T$ increased and calculated performance metrics over all states and time periods, as in Figure 2, I would be unable to know whether changes in accuracy or efficiency were due to changing model performance or changing composition of the test set. I fit these sets of models to simulated data with low ($\psi, \omega = 1$) and high ($\psi, \omega = 100$) temporal stability.

Figure 3 displays how bias, variance, and RMSE change as the number of time periods seen by the model increases. As in Figure 2, trend lines are plotted using LOESS and give 95 percent confidence intervals. Estimates from no-pooling and moving average models do not change as the number of time periods increases, so they appear as horizontal lines.[14]

There is little evidence of trends in bias, save for the very wide confidence intervals on moving average, linear trend, and year-intercepts models under conditions of low opinion stability. This suggests these models are producing inaccurate estimates with some observations' estimates biased upward and others downward, averaging each other out in the aggregate. Increasing RMSE—partially comprised of bias—for the latter two models corroborates this interpretation. Under conditions of high opinion stability, scholars can decrease bias in models with demographic-year random intercepts and local-level transitions by adding more time periods, but there is little marginal gain after four or five periods.

The more instructive insights in this analysis come from the variance and RMSE plots. Overall, the variance of each estimator decreases as the number of time periods increases. That is, as the dynamic models see data from a broader time horizon, their estimates of state-level opinion at $t_1$ tend to become more efficient. This is especially true when opinion is highly stable—the less

[14]The no-pooling model is fit only on $t_1$ and the moving average model is fit with the first three years of data.

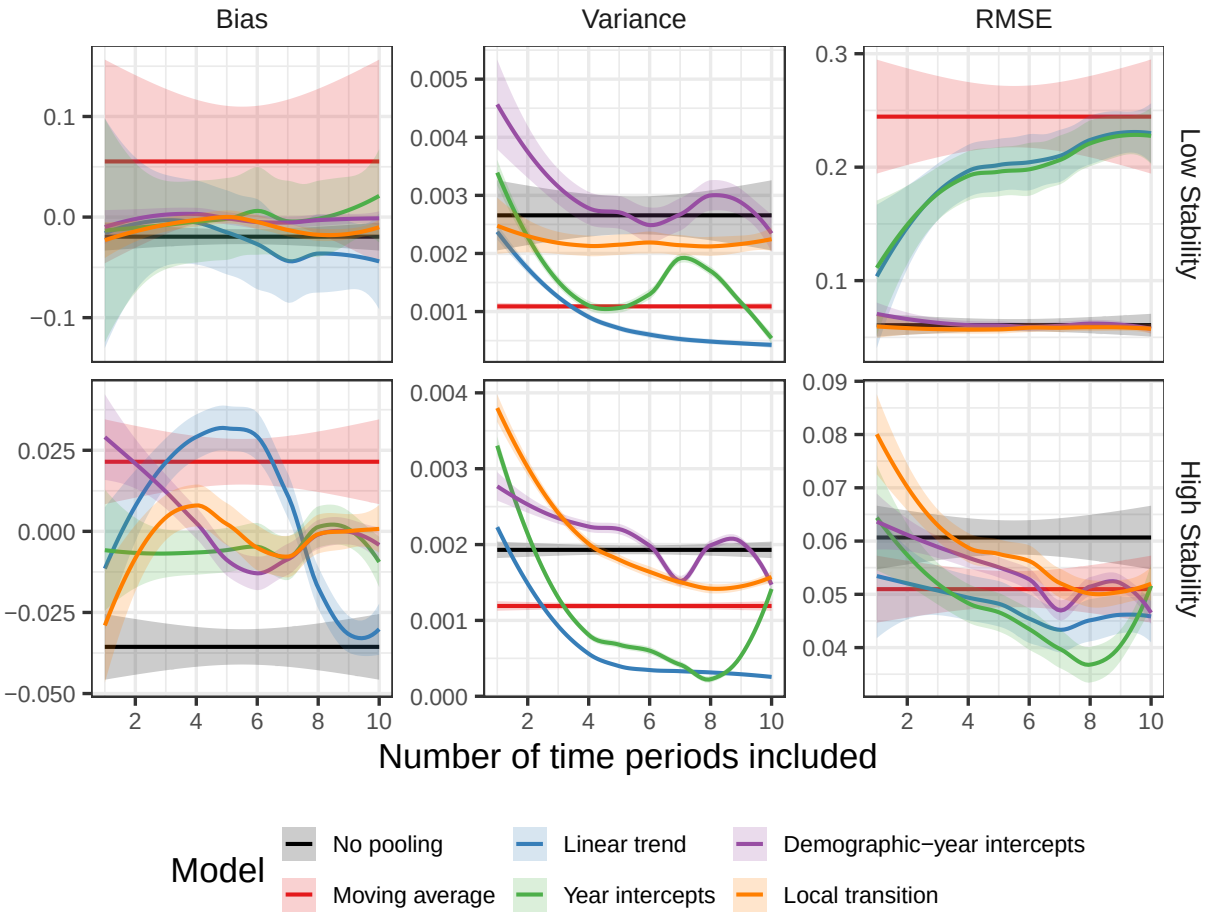


Figure 3: Model Performance at $t_1$ as $T$ Increases. Performance metrics reflect model estimates for all states at $t_1$ within each iteration. Trend lines show 95% confidence intervals.

opinion fluctuates, the more informative each additional time period is to estimating opinion in the target time period. In fact, at high levels of stability, all dynamic models—plus the moving average model—converge to approximately the same performance as $T$ increases. When the number of time periods is few, the local-level transition model is still by far the most error-prone, owing both to its relatively inaccurate estimates and the high uncertainty with which they are calculated. When opinion is highly volatile, the no-pooling, demographic-year random intercepts, and local-level transition models all exhibit equally strong performance, while the moving average model is uniquely poor.

These results suggest three conclusions. First, the highly variable performance seen among CES policy preferences may be partially explained by the variation in time series length, at least for a couple models. Second, unless opinion fluctuates wildly over time, adding more time periods may result in more efficient estimates with little, if any, loss of accuracy. Third, although it displays higher variance in its estimates relative to several other models, the model with demographic-year intercepts appears the most versatile, exhibiting low RMSE regardless of the level of opinion stability or number of time periods included in the model.

## Recovering State-Level Opinion with Scarce Data

Since one of MRP's main benefits is helping correct for undesirable finite-sample properties, past authors have frequently pointed to state population and sample sizes as important determinants in model performance (Bisbee, 2019; Lax & Phillips, 2009b; Ornstein, 2020). In the dynamic context, it is not only sample size across *states* that matters, but also across *time periods*.

To evaluate the sensitivity of dynamic MRP to sample size constraints, I again examine model performance at $t_1$. This time, however, I include all time periods in the model and vary the sample size at $t_1$, holding sample sizes at all other time periods $t \in \{2, ..., T\}$ constant. As in the analysis of increasing time series length, I again evaluate model performance in low- and high-stability scenarios. This simulation allows me to assess how dynamic MRP performs when observations are unevenly distributed among time periods, similar to analyses of small- and large-state performance in static MRP; how robust dynamic MRP is to small samples; and whether researchers may be able to leverage information from other time periods to improve inference in cases where data is scarce.

Figure 4 shows how performance metrics change as sample size increases. The x-axis depicts the ratio of $t_1$ sample size to the sample size at all other time periods, the latter of which is held constant at 1,000 observations. The lowest end of the x-axis therefore reflects a sample size of only 100 total observations at $t_1$ (ten per state, on average), while the upper end reflects a sample size on par with all other time periods.

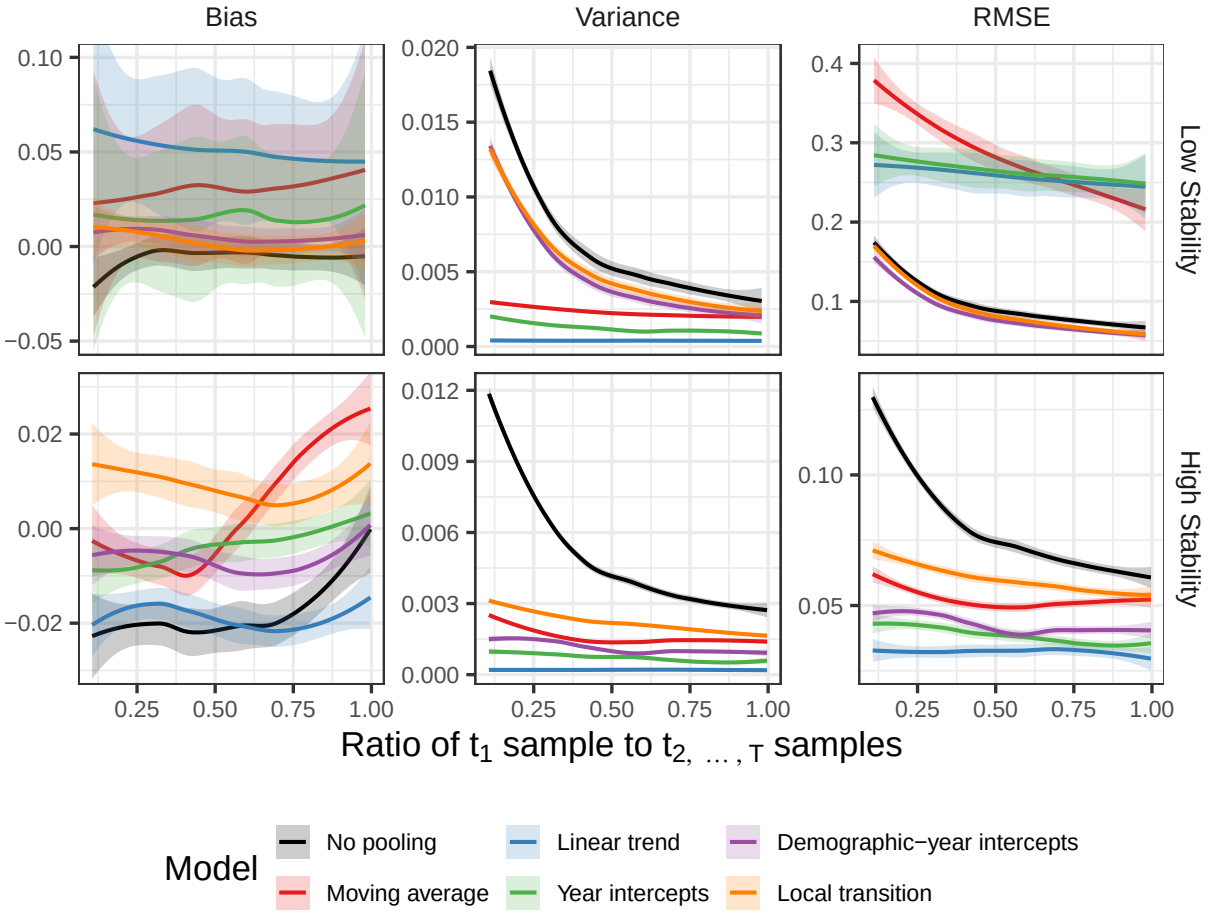


Figure 4: Model Performance at $t_1$ as Sample Size Increases. Performance metrics reflect model estimates for all states at $t_1$ within each iteration. Trend lines show 95% confidence intervals.

Results again show an interesting interaction between sample size and opinion stability. When opinion changes rapidly over time, demographic-year random intercepts and local-level transition models again show relatively high variance in their estimates, but this variance decreases as sample size increases and is completely washed out in RMSE by high bias in other models' estimates. At high volatility, no-pooling performs nearly on par with these more complex models.

When opinion is very stable over time, however, a no-pooling model is decidedly suboptimal. It returns highly biased and variable estimates when sample sizes are small. Both metrics improve as sample size increases, but never match the performance gained from even the worst dynamic model. Somewhat surprisingly, the local-level transition model also struggles to recover accurate

estimates when sample sizes are low; combining this comparatively high bias with slightly higher variance makes this model a notably worse option than other dynamic models. By contrast, the linear trend model performs well in the high-stability scenario. Although its estimates are less accurate than other models', the precision with which they are estimated leads to favorable RMSE metrics. This may be partially attributable to the data-generating process, as a more stable time trend gives the model more structure on which to fit.

As in the analysis of increasing $T$ in Figure 3, the model with demographic-year random intercepts is the most versatile in this simulation. Though it has relatively high variance in low-stability situations, this variance decays quickly at higher sample sizes, and its near-zero bias produces excellent overall performance. Judging by the summary RMSE metric, this model can estimate subnational opinion with little to no dropoff in performance. As shown in Figure 3, this performance can be achieved with only a few additional time periods of data. Even if researchers are not interested in temporal trends, they can nevertheless use data from other time periods to improve cross-sectional opinion estimates in the time periods they do require.

# Discussion and Recommendations

There are many seemingly reasonable approaches to incorporating time into subnational opinion estimation frameworks. I investigated six such approaches here, but scholars should continue to test whether adjustments to these general specifications can improve accuracy in each unique use case. Evidence from public opinion on a wide range of policy issues suggests that no one model is universally desirable; models perform well on some issues and poorly on others. When possible, applied researchers using dynamic MRP should consider testing several appropriate model specifications and reporting the results of validation exercises.

In the common scenario where validation is not feasible, however, Monte Carlo simulations can provide some valuable guidance on model selection. Much of the variability in model performance exhibited in the CES data, I suggest, is attributable to characteristics of the time series themselves:

over-time opinion volatility, time series length, and sample size. Particularly important for model selection is these factors' differential effects on bias and variance.

The no-pooling model, popular among applied researchers, exhibits the highest degree of uncertainty in all scenarios. Its estimates are often more *accurate* than other models, but they are so inefficiently estimated that the no-pooling model is rarely a good choice and frequently the worst possible one. Especially if researchers wish to use these estimates in downstream analyses, where the measurement error should be propagated through to correlation or effect estimates, they would be better off employing a more complex model that can draw on information from multiple time periods at once.[15] The moving-average model is an equally simple, entry-level model that incorporates more information, but it often returns estimates with such bias that it should only be used if the time series is known to be extremely volatile.

In Gelman et al.'s (2018) analyses of same-sex marriage opinion, they find models with linear time trends frequently best the others. Across twenty-nine issues and several simulations, I replicate this finding only for a small handful of issues and a couple simulated scenarios. The same is true for a model with random intercepts by year. For example, when opinion is highly stable over time, these models do indeed provide reasonable options that can often return accurate, highly efficient estimates. In fact, when opinion is highly stable and there are many time periods available, scholars should consider using a model with random intercepts by year, as it provides low bias without the higher variance exhibited by more complex models. However, performance is on a knife's edge; small increases in opinion fluctuation over time or decreases in the number of time periods available to the model can lead to a precipitous decline in performance. For this reason, researchers may wish to reach instead for a model that is more versatile, if slightly less efficient.

Linear trend and year-intercept models' struggles suggest that what matters more than accounting for time trends in the outcome variable is accounting for time trends in the *relationships* of demographic predictors to the outcome variable. Demographic-year intercept and local-level transition models achieve this, which may partially explain why they perform well in many different

[15] I focus on policy responsiveness to same-sex marriage opinion in SI section D, demonstrating that substantive inferences can change based on the model selected.

scenarios. Local-level transition models are clearly suboptimal when opinion is stable and there are few time periods, but they are on par with or better than other options in most other situations.

Across all simulations, the demographic-year random intercepts model is clearly the most flexible of the models tested. It does not always have the lowest bias or highest efficiency, but it is never far from the model that does, and it is highly performant in an impressively wide range of scenarios. Averaged across all CES issues, it returns the second-lowest RMSE, just behind models with random intercepts by year. If scholars seek a model that is a safe bet no matter the application, this one would be a defensible choice.

Most work using MRP applies the method to binary outcome variables—such as two-party presidential vote or policy support (Enns & Koch, 2013; Kastellec, 2018; Kuriwaki et al., 2024)—even if that requires recoding survey items originally asked with more than two response options. Accordingly, I formulated all candidate models for binary outcome variables. But these models can also be extended to other types of data structures. In SI section E, I replicate a study of state-level racial resentment over time by Smith et al. (2020), where the outcome variable—an index derived from the racial resentment survey battery—is continuous. Because MRP is based on a generalized linear model, it can connect to nearly any data source that can be modeled with a distribution from the exponential family.

Similarly, I focused on public opinion because it is the most common use case for MRP. But MRP is part of a broader ecosystem of methods for small-area estimation, applied to diverse topics such as urban planning, literacy, and agriculture (Kontokosta et al., 2018; Pfefferman et al., 2008; Singh et al., 2002). Future work should assess the appropriateness of dynamic MRP for these related fields, relative to other statistical approaches for incorporating time into small-area estimation (Rao & Yu, 1994; Singh et al., 2005).

# References


Arceneaux, K. (2001). The "Gender Gap" in State Legislative Representation: New Data to Tackle an Old Question. *Political Research Quarterly*, *54*(1), 143–160.

Ben-Shalom, Y., Martinez, I., & Finucane, M. M. (2021). Risk of Workforce Exit due to Disability: State Differences in 2003–2016. *Journal of Survey Statistics and Methodology*, *9*(2), 209–230.

Bisbee, J. (2019). BARP: Improving Mister P Using Bayesian Additive Regression Trees. *American Political Science Review*, *113*(4), 1060–1065.

Blackwell, M. (2013). A Framework for Dynamic Causal Inference in Political Science. *American Journal of Political Science*, *57*(2), 504–520.

Brace, P., Sims-Butler, K., Arceneaux, K., & Johnson, M. (2002). Public Opinion in the American States: New Perspectives Using National Survey Data. *American Journal of Political Science*, *46*(1), 173.

Broniecki, P., Leemann, L., & Wüest, R. (2022). Improved Multilevel Regression with Poststratification through Machine Learning (autoMrP). *The Journal of Politics*, *84*(1), 597–601.

Buttice, M. K., & Highton, B. (2013). How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys? *Political Analysis*, *21*(4), 449–467.

Butz, A. M., & Kehrberg, J. E. (2016). Estimating Anti-Immigrant Sentiment for the American States using Multi-Level Modeling and Post-Stratification, 2004–2008. *Research & Politics*, *3*(2), 205316801664583.

Caughey, D., & Warshaw, C. (2015). Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model. *Political Analysis*, *23*(2), 197–211.

Caughey, D., & Warshaw, C. (2018). Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014. *American Political Science Review*, *112*(2), 249–266.

Caughey, D., & Warshaw, C. (2019). Public Opinion in Subnational Politics. *The Journal of Politics*, *81*(1), 352–363.

Claassen, C., & Traunmüller, R. (2020). Improving and Validating Survey Estimates of Religious Demography Using Bayesian Multilevel Models and Poststratification. *Sociological Methods & Research*, *49*(3), 603–636.

Clark, A. K. (2017). Updating the Gender Gap(s): A Multilevel Approach to What Underpins Changing Cultural Attitudes. *Politics & Gender*, *13*(01), 26–56.

Enns, P. K., & Koch, J. (2013). Public Opinion in the U.S. States: 1956 to 2010. *State Politics & Policy Quarterly*, *13*(3), 349–372.

Erikson, R. S., Wright, G. C., Jr., & McIver, J. P. (1993). *Statehouse Democracy: Public Opinion and the American States*. Cambridge University Press.

Franko, W. W. (2017). Understanding Public Perceptions of Growing Economic Inequality. *State Politics & Policy Quarterly*, *17*(3), 319–348.

Gao, Y., Kennedy, L., Simpson, D., & Gelman, A. (2021). Improving Multilevel Regression and Poststratification with Structured Priors. *Bayesian Analysis*, *16*(3).

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Lax, J., Phillips, J., Gabry, J., & Trangucci, R. (2018, August). *Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion* [Working Paper]. http://www.stat.columbia.edu/~gelman/research/unpublished/MRT(1).pdf

Gelman, A., & Little, T. C. (1997). Poststratification into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*, *23*(2), 127–135.

Giraudy, A. (2015). *Democrats and Autocrats: Pathways of Subnational Undemocratic Regime Continuity within Democratic Countries*. Oxford University Press.

Giraudy, A., Moncada, E., & Snyder, R. (2019). Subnational Research in Comparative Politics: Achievements and Future Directions. In A. Giraudy, E. Moncada, & R. Snyder (Eds.), *Inside Countries: Subnational Research in Comparative Politics*. Cambridge University Press.

Goplerud, M. (2024). Re-Evaluating Machine Learning for MRP Given the Comparable Performance of (Deep) Hierarchical Models. *American Political Science Review*, *118*(1), 529–536.

Grumbach, J. M. (2022). *Laboratories Against Democracy: How National Parties Transformed State Politics*. Princeton University Press.

Kastellec, J. P. (2018). Judicial Federalism and Representation. *Journal of Law and Courts*, *6*(1), 51–92.

Key, V. (1949). *Southern Politics in State and Nation*. Alfred A. Knopf.

Knox, D., Lucas, C., & Cho, W. K. T. (2022). Testing Causal Theories with Learned Proxies. *Annual Review of Political Science*, *25*, 419–441.

Kołczynska, M., Bürkner, P.-C., Kennedy, L., & Vehtari, A. (2024). Modeling Public Opinion over Time and Space: Trust in State Institutions in Europe, 1989-2019. *Survey Research Methods*, *18*(1), 1–19.

Kontokosta, C. E., Hong, B., Johnson, N. E., & Starobin, D. (2018). Using Machine Learning and Small Area Estimation to Predict Building-Level Municipal Solid Waste Generation in Cities. *Computers, Environment and Urban Systems*, *70*, 151–162.

Kuriwaki, S., Ansolabehere, S., Dagonel, A., & Yamauchi, S. (2024). The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level. *American Political Science Review*, *118*(2), 922–939.

Lax, J. R., & Phillips, J. H. (2009a). Gay Rights in the States: Public Opinion and Policy Responsiveness. *American Political Science Review*, *103*(3), 367–386.

Lax, J. R., & Phillips, J. H. (2009b). How Should We Estimate Public Opinion in the States? *American Journal of Political Science*, *53*(1), 107–121.

Leemann, L., & Wasserfallen, F. (2017). Extending the Use and Prediction Precision of Subnational Public Opinion Estimation. *American Journal of Political Science*, *61*(4), 1003–1022.

Lewis, D. C., & Jacobsmeier, M. L. (2017). Evaluating Policy Representation with Dynamic MRP Estimates: Direct Democracy and Same-Sex Relationship Policies in the United States. *State Politics & Policy Quarterly*, *17*(4), 441–464.

Mehlhaff, I. D. (forthcoming). *Mass Polarization across Time and Space*. Cambridge University Press.

Ornstein, J. T. (2020). Stacked Regression and Poststratification. *Political Analysis*, *28*(2), 293–301.

Pacheco, J. (2011). Using National Surveys to Measure Dynamic U.S. State Public Opinion: A Guideline for Scholars and an Application. *State Politics & Policy Quarterly*, *11*(4), 415–439.

Pacheco, J. (2014). Measuring and Evaluating Changes in State Opinion Across Eight Issues. *American Politics Research*, *42*(6), 986–1009.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*, *12*(4), 375–385.

Pfefferman, D., Terryn, B., & Moura, F. A. (2008). Small Area Estimation under a Two-Part Random Effects Model with Application to Estimation of Literacy in Developing Countries. *Survey Methodology*, *34*(2), 235–249.

Rao, J. N. K., & Yu, M. (1994). Small-Area Estimation by Combining Time-Series and Cross-Sectional Data. *Canadian Journal of Statistics*, *22*(4), 511–528.

Ruggles, S., Flood, S., Sobek, M., Backman, D., Chen, A., Cooper, G., Richards, S., Rodgers, R., & Schouweiler, M. (2024). *IPUMS USA* (Version 15). Minneapolis, MN.

Sellers, J. M. (2019). From Within to Between Nations: Subnational Comparison across Borders. *Perspectives on Politics*, *17*(1), 85–105.

Shirley, K. E., & Gelman, A. (2015). Hierarchical Models for Estimating State and Demographic Trends in US Death Penalty Public Opinion. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *178*(1), 1–28.

Simonovits, G., & Bor, A. (2023). Stability and Change in the Opinion–Policy Relationship: Evidence from Minimum Wage Laws. *Research & Politics*, *10*(3), 1–7.

Singh, B. B., Shukla, G. K., & Kundu, D. (2005). Spatio-Temporal Models in Small Area Estimation. *Survey Methodology*, *31*(2), 183–195.

Singh, R., Semwal, D. P., Rai, A., & Chhikara, R. S. (2002). Small Area Estimation of Crop Yield using Remote Sensing Satellite Data. *International Journal of Remote Sensing*, *23*(1), 49–56.

Smith, C. W., Kreitzer, R. J., & Suo, F. (2020). The Dynamics of Racial Resentment across the 50 US States. *Perspectives on Politics*, *18*(2), 527–538.

Tai, Y., Hu, Y., & Solt, F. (forthcoming). Democracy, Public Support, and Measurement Uncertainty. *American Political Science Review*.

Warshaw, C., & Rodden, J. (2012). How Should We Measure District-Level Public Opinion on Individual Issues? *The Journal of Politics*, *74*(1), 203–219.

Wiertz, D., & Lim, C. (2021). The Rise of the Nones across the United States, 1973 to 2018: State-Level Trends of Religious Affiliation and Participation in the General Social Survey. *Sociological Science*, *8*, 429–454.