Improving Subnational Opinion Estimation from Cluster-Sampled Polls*

Michael Auslen†

March 19, 2024

Abstract

The development of Multilevel Regression and Poststratification (MRP) has allowed scholars to more accurately estimate subnational public opinion using national polls. However, MRP generally recovers less accurate estimates from polls whose respondents are selected using cluster sampling—also called area-probability sampling. This is in part because cluster-sampled polls rely on a complex form of random sampling focused on national representativeness that may result in small or unrepresentative subsamples in subnational geographies. This has limited MRP's usefulness in subnational opinion estimation in several contexts, including historical polls in the United States, where cluster-sampling was common into the 1980s, and large academic studies in many countries today. In this paper, I propose two approaches to improve estimation from MRP with cluster-sampled polls. The first is pooling data from multiple surveys to produce a larger sample of clusters. The second is Clustered MRP (CMRP), which extends MRP by modeling opinion using the geographic information included in a survey's cluster-sampling procedure. Using simulations, I show that both methods improve upon traditional MRP, and I validate them using historical polls in the United States.

Words: 8,681

*I thank Naoki Egami, Andrew Gelman, Max Goplerud, Justin Phillips, Robert Shapiro, Elizabeth Zechmeister, and participants at APSA 2021, the 2022 State Politics and Policy Conference, and the Columbia American Politics Graduate Student Workshop for helpful feedback. I also thank Kathleen Weldon from the Roper Center for Public Opinion Research for sharing internal memos from the Gallup Poll. I acknowledge computing resources from Columbia University's Shared Research Computing Facility project (supported by NIH Research Facility Improvement Grant 1G20RR030893-01 and New York State Empire State Development, Division of Science Technology and Innovation Contract C090171).

†Ph.D. candidate, Department of Political Science, Columbia University, New York, NY. Email: michael.auslen@columbia.edu. URL: `https://michaelauslen.com`.

Estimates of subnational public opinion are necessary to study many important questions in political science. Despite their usefulness, accurate estimates can be difficult to obtain due to cost of obtaining sufficiently many samples in all subnational units of interests (e.g., states or regions). In response to this lack of data, scholars have for decades developed alternative approaches to estimate subnational opinion from national polls (e.g., Erikson, Wright and McIver 1993).

In recent years, subnational opinion estimation has been substantially aided by the development of Multilevel Regression and Poststratification (MRP), which allows for more accurate estimates than previous methods (Park, Gelman and Bafumi 2004). MRP involves fitting a predictive model of individual opinion from survey data, predicting opinion for different demographic and geographic subgroups in the public, and then taking a weighted average using the known distribution of these subgroups within subnational geographies. It has become the gold standard for estimating opinion, primarily based on studies of state and legislative district opinion in the United States (Lax and Phillips 2009*b*; Warshaw and Rodden 2012) and in Europe (Leemann and Wasserfallen 2016; Toshkov 2015; Lipps and Schraff 2021).

The development of MRP has made it considerably more straightforward for scholars to estimate opinion on individual issues within subnational geographies. This has proven particularly useful for studies of elite responsiveness (Lax and Phillips 2009*a*, 2012; Tausanovitch and Warshaw 2014), electoral behavior (Ghitza and Gelman 2013; Gelman et al. 2016), and public opinion (Shirley and Gelman 2015), among others. By estimating opinion on individual issues, scholars can examine more specific determinants of policy and opinion than latent opinion measures allow.

While MRP has provided considerable advancements, it has primarily been developed with polls whose respondents represent simple random samples (or close approximations of simple random samples) of the target population in mind. Yet, there are many settings in which such samples cannot be produced. In such cases, scholars have instead turned to

alternative sampling designs. One popular alternative is cluster sampling—also called area-probability sampling—which produces a sample by taking multiple respondents from a small number of randomly drawn geographic areas.

Cluster sampling was nearly ubiquitous among U.S. polling firms into the 1980s, as it allowed for high-quality national samples to be obtained for face-to-face interviews at a comparatively low cost (Warshaw 2016). It is also a common sampling method for large, multinational and national academic surveys (e.g., AmericasBarometer, Asian Barometer, the American National Election Study [ANES], and General Social Survey [GSS]). As a result, public opinion data in many contexts has been predominantly generated through the use of cluster-sampling methods. These procedures commonly draw clusters nationally or within large regions, conditional on some geographic characteristics (e.g., the "urbanness" of communities). This allows them to produce *nationally* representative samples in a single poll without necessarily generating representative samples at *subnational* levels of interest, such as states. This makes these polls particularly susceptible to inaccurate opinion estimation at smaller geographies using conventional MRP approaches (Stollwerk 2017).

As a result, scholarship is limited in any domain which relies on subnational estimates of opinion. Among these are descriptive studies of opinion across space and time, as well as research on policy responsiveness, which uses opinion as an independent variable to predict the positions taken by legislators, governments, or political parties. Measurement error in the opinion variable may introduce a number of challenges, including the attenuation of regression coefficients. And common workarounds, such as pooling and disaggregating surveys over many years (e.g., a decade or more) do not allow for the estimation of effects over time.

In this paper, I introduce and test two approaches that scholars may employ to improve opinion estimation with MRP from cluster-sampled polls. I begin by providing background information about the MRP and cluster sampling methods that underpin the following sections. In Section 2, I illustrate problems that may arise from using traditional MRP in the

case of cluster-sampled polls. I first use an empirical example of abortion opinion in 1980, in which MRP returns estimates that lack face validity. I show that the uneven distribution of clusters across states is likely the source of the problem. I clarify this intuition using simulations that demonstrate the limitations of traditional MRP under this sampling design.

Sections 3 and 4 propose two possibly complementary solutions to the problem. In each section, I conduct a simulation analysis and validate the methods against the case of presidential polls. My first solution, outlined in Section 3, pools responses from multiple surveys. This can address the problem by increasing the number of distinct clusters and thus hopefully mitigating problems inherent to a single survey. However, there are significant data limitations that may make this approach impossible in many circumstances. First, few issues are repeatedly surveyed using the same question wording over narrow windows of time. Second, I find that pooling only produces significant improvements in estimation when the number of clusters (and not merely sample size within clusters) increases. As a result, it would generally be necessary to find the same question asked in polls fielded by different firms.

Section 4 presents an alternative, model-based strategy—Clustered MRP (CMRP)—that incorporates features of sampling design into the model. By explicitly including the geographic levels used in a pollster's sampling procedure in MRP's predictive model, CMRP properly accounts for polling firms' sampling protocols and allows information from similar geographic areas to be pooled across state and regional lines. Even using a single poll, I find that CMRP can reduce mean absolute error (MAE) in state-level opinion estimates by 2.5-4% compared to standard MRP approaches. These accuracy gains are similar in magnitude to those associated with other improvements to MRP, such as using machine learning or models with deeper interactions for opinion estimation with modern polls (Ornstein 2020; Goplerud 2023).

Finally, I also discuss the concerns of particular cluster-sampling procedures and steps needed to produce poststratification data from the Census for CMRP in Section 5, as well as

in the appendix. While this paper focuses on historical polls in the U.S., my approach may be applicable in other contexts in which cluster-sampling is common, including comparative multinational surveys.

1 MRP and Cluster Sampling

In recent decades, scholars have turned to MRP to estimate subnational public opinion from the individual responses in national polls. Developed by Gelman and Little (1997) and Park, Gelman and Bafumi (2004), MRP has been shown to outperform alternative methods, such as disaggregation, at the state (Lax and Phillips 2009*b*), congressional and state legislative district (Warshaw and Rodden 2012), and municipal levels (Tausanovitch and Warshaw 2013). This is because MRP uses both demographic and geographic characteristics of respondents and partially pools information across geographies to better model the drivers of individual opinion.

Particularly compelling for applied researchers, MRP has the potential to produce reliable estimates of state opinion from a single, conventional national survey (Lax and Phillips 2009*b*), although it may produce less accurate estimates from such polls when geographic variables are poor predictors of individual-level opinion (Buttice and Highton 2013).

1.1 MRP with Polls using Simple Random Samples

There are two steps to estimating opinion with MRP: First, a multilevel model is fit to predict individual response to a binary question using individual-level data from a survey. Then, using the model and the joint distribution of demographic characteristics in the population from the Census, the scholar poststratifies to estimate the average response to the question of interest at the state (or other subnational) level.

The typical multilevel logistic model for MRP contains as predictors detailed demographic information about respondents, state and region indicators, and one or two state-level vari-

ables (e.g., presidential vote and religiosity). Below, I formalize a standard MRP model that could be fit from data available in a standard Gallup Poll fielded in the U.S. during the 1980s, using the notation from Gelman and Hill (2007). The outcome of interest, y_i, indicates an individual respondent i's response to a survey question. The α terms are random effects corresponding to demographic or geographic groups; so $\alpha^{\text{race}}_{r[i]}$ indicates the random effect for the racial group r to which respondent i belongs. The state random effect $\alpha^{\text{state}}_{s[i]}$ is modeled as a function of the region and political variables included (here I use $RepVote$, Republican vote share in the last presidential election, and $Relig$, the proportion of the state that identifies as evangelical Christian or Mormon).

$$
\begin{aligned}
\Pr(y_i = 1) &= \text{logit}^{-1}(\beta^0 + \alpha^{\text{race}}_{r[i]} + \alpha^{\text{sex}}_{g[i]} + \alpha^{\text{age}}_{k[i]} + \alpha^{\text{educ}}_{l[i]} + \alpha^{\text{state}}_{s[i]}) \\
\alpha^{\text{race}}_{r} &\sim N(0, \sigma^2_{race}), \text{ for } r = 1, ..., 3 \\
\alpha^{\text{sex}}_{g} &\sim N(0, \sigma^2_{sex}), \text{ for } g = 1, 2 \\
\alpha^{\text{age}}_{k} &\sim N(0, \sigma^2_{age}), \text{ for } k = 1, ..., 4 \\
\alpha^{\text{educ}}_{l} &\sim N(0, \sigma^2_{educ}), \text{ for } l = 1, ..., 4 \\
\alpha^{\text{state}}_{s} &\sim N(\alpha^{\text{region}}_{m[s]} + \beta^{\text{RepVote}} * RepVote_s + \beta^{\text{Relig}} * Relig_s, \sigma^2_{state}), \text{ for } s = 1, ..., 50 \\
\alpha^{\text{region}}_{m} &\sim N(0, \sigma^2_{region}), \text{ for } m = 1, ..., 8
\end{aligned}
\tag{1}
$$

This model allows us to predict the expected level of support for the policy y among each "type" of person in the population—that is, each of the 4,800 possible combinations of $race \times sex \times age \times educ \times state$. These predictions are then used to poststratify and aggregate to the state level by taking a weighted average where the weights are the share of each combination of demographic variables in the state's population.

Estimates may be further improved by fitting a "deep" model with interactions among demographic and geographic variables in ways not captured by a simple model without interactions (Ghitza and Gelman 2013; Goplerud 2023).

1.2 Cluster-Sampled Surveys

The above model has been developed assuming survey respondents are independently drawn from the population, as in a simple random sample. However, this assumption may not hold in many circumstances (Berinsky 2017). Instead, when such sampling methods are impossible or impractical, pollsters often rely on alternative methods to produce samples that are representative of target populations. For many surveys, this target is the population of an entire country, and not any one subnational unit.

One of the most common procedures used for survey sampling is cluster sampling (also called area-probability sampling). Variations of this approach (and especially multi-stage sampling methods) were almost universally used by U.S. polling firms from the 1950s–1980s (Warshaw 2016).[1] Two major academic studies, the GSS and ANES, still use cluster-sampling to produce all or part of their samples today, as do many large multinational surveys (see, e.g., Latin American Pulic Opinion Project 2019; Asian Barometer Survey 2003). As a result, the only quality polls available to scholars in many contexts are likely to be cluster-sampled.

The aim of cluster sampling is straightforward: researchers randomly select a set of "clusters," such as cities or neighborhoods, from which they randomly draw people to interview. This approach is appealing to survey researchers because it reduces the costs of producing a nationally representative sample for surveys that rely on in-person interviews.

As an example, consider the Gallup Poll during the late 1970s and early 1980s, on which I based the sampling algorithm in the simulation studies discussed below (Gallup Organization 1980*b*). First, Gallup assigned each state to a region. Within each region, geographic areas were assigned to a "size-of-community stratum" based on urban/rural status and population. These region-stratum combinations form the basis of primary sampling units (PSUs). Then, Gallup randomly selected two localities from each PSU, weighting by population, and

[1] By the 1988 election, major pollsters used random-digit dialing to produce samples (Voss, Gelman and King 1995).

repeated the process using progressively smaller geographies to identify a block or cluster of blocks. The resulting sample is expected to produce reasonable estimates of national opinion. A more detailed description of this procedure can be found in Appendix A.

2 Challenges of MRP with Cluster-Sampled Polls

While some scholars have used MRP with cluster-sampled polls (e.g., Shirley and Gelman 2015), its performance has been primarily validated on polls that use simple random samples (or close approximations). As a result, it is less clear whether MRP should perform well under more complex sampling procedures—such as cluster sampling—in which respondents are not drawn independently from the public (Stollwerk 2017). In this section, I provide a simple empirical example that illustrates problems that may arise.

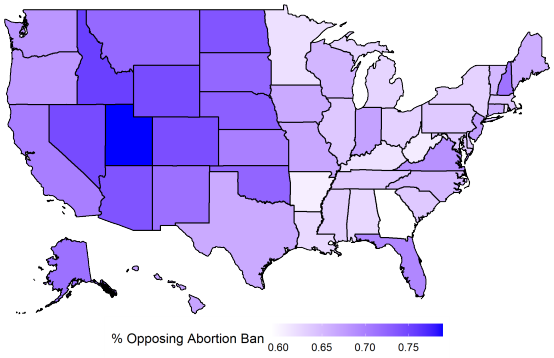
2.1 Case Study: Abortion Opinion

To illustrate the problems that may arise from using MRP with cluster-sampled polls, I estimated opinion on abortion from a survey fielded by Gallup in September 1980 and downloaded from the Roper Center for Public Opinion Research (Gallup Organization 1980*a*). The survey is typical of the period and used cluster sampling to produce a set of respondents ($N = 1,602$) who were interviewed in person at their homes. I used MRP to estimate state-level opinion using responses to a question asking whether respondents "generally favor" or "generally oppose" an ban on abortion.[2]

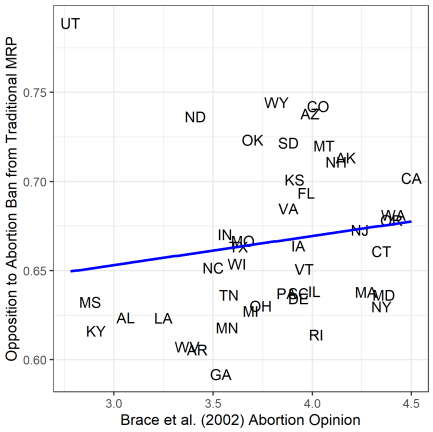
I modeled individual opinion in Stan (Carpenter et al. 2017) with random effects for race, female, the race×female interaction, age group, education, state, and region. I also included state Republican vote share and the share of the population that is evangelical Christian or Mormon as linear predictors. This specification is typical for MRP for a question about social issues like abortion. All variables in the model were used in poststratification, using weights

[2]The exact wording of the question can be found in Appendix F.

Figure 1: Opposition to Abortion Ban from Traditional MRP on 1980 Gallup Poll



(a) State-Level Opinion



(b) Correlation with Abortion Opinion

Note: The lefthand panel plots opposition to a ban on abortions estimated using Traditional MRP. Darker states are more opposed to a ban. The righthand panel plots estimated opposition against state estimates of abortion liberalism from Brace et al. (2002). The blue curve is a least-squares regression of the relationship between the variables.

from joint distributions of the population downloaded from IPUMS-NHGIS (Manson et al. 2021). Presidential vote data are from Leip (2021), and religion data are from the Churches and Church Membership in the United States study (Grammich et al. 2019).

Figure 1 reports the results from the MRP model. The lefthand panel maps the share of each state that opposed a ban on abortion, as estimated with MRP. The righthand panel compares modeled opposition to an abortion ban with Brace, Sims-Butler, Arceneaux and Johnson's (2002) measure of state-level abortion liberalism, which they produced by pooling surveys from 1974–1998.[3] A higher score on their scale corresponds to more liberal opinion on abortion. The MRP results are clearly only minimally correlated with the baseline estimates of abortion opinion. Similar analyses using latent policy liberalism estimates provide similar results (these can be found in Appendix F). Many state estimates also clearly lack face validity. For example, it is unlikely that Utah would have the highest opposition to a ban on abortions in the country. Conversely, more liberal states in the Northeast and upper

[3]Due to data limitations, Brace et al. (2002) do not publish opinion estimates for Hawaii, Idaho, Maine, Nebraska, Nevada, and New Mexico. I likewise drop these states from the plot in Figure 1b.

Table 1: Utah Respondents in Gallup Abortion Poll

	Opinion	**Race**	**Sex**	**Age Group**	**Education**	**City Size**	**Party**
1	(No response)	White	Female	18-24	no hs	50,000-99,999 & sub.	Rep.
2	(No response)	White	Female	65+	college	50,000-99,999 & sub.	Dem.
3	Support	White	Male	45-64	some college	50,000-99,999 & sub.	Dem.
4	Support	White	Male	45-64	some college	50,000-99,999 & sub.	Ind.
5	(No response)	White	Male	25-44	high school	50,000-99,999 & sub.	Ind.

Midwest show surprisingly low opposition to banning abortions.

One reason for error is that state estimates may depend only on a small number of clusters from particular (non-representative) parts of the state. For example, Table 1 shows the five Utah respondents in the poll. Of the five, which all appear to come from one cluster in a city of 50,000 to 99,999 people, only two answered the question about an abortion ban, and both were opposed.[4] Respondents in the Utah cluster are more likely to be Democrats or Independents than Utahans of the era as a whole (Erikson, Wright and McIver 1993). In principle, the limited nature of this Utah subsample is the type of problem MRP handles well. However, in the case of cluster sampling, the predictive model attributes the average opinion in this cluster to Utah as a whole, rather than the stratum×region combination that it was drawn into the sample to represent.

Although subgroup means converge on true population means when a large number of clusters are included in a sample (Kish and Frankel 1974), scholars often have access to a limited number of surveys (indeed, there may be only one survey asking a particular question in the time period of interest). As a result, the relevant question for using MRP with cluster-sampled polls not whether state-level subsamples are representative of the population in expectation over repeated surveys, but rather whether the respondents from a given state *in a single poll* are likely to be representative. In Appendix B, I show that state subsamples in individual polls are often not representative of their states. This problem is especially severe in states with lower populations, which are much more likely to have no respondents included

[4]In the poll nationwide, just 4.2% of respondents did not respond to the question.

in a poll, and whose subsamples are less representative of the state population (because they include fewer clusters).[5] It is also intuitively likely to be the case in states with more diverse and segregated populations, where any two clusters may be very different from one another. Erikson, Wright and McIver (1993) note these potential problems in discussing their decision to use disaggregated CBS/*New York Times* polls, rather than a cluster-sampled survey.

2.2 Simulation Study 1: MRP with Cluster-Sampled Polls

To better understand and illustrate the problems that may arise when using MRP with cluster-sampled polls, I conducted a series of simulations.

In each simulation, I generated 1 million "voters" distributed across 50 states (according to their actual share of the population) and seven size-of-community strata (according to a random distribution that I hold constant across simulations). Next, I randomly assigned each voter a binary demographic predictor (which I call race) distributed according to each state's real-world White and non-White populations. I then drew a survey response for each voter such that:

$$
\begin{aligned}
y_i &\sim \text{Bern}(\text{logit}^{-1}(\alpha^{\text{race}}_{r[i]} + \alpha^{\text{state}}_{s[i]} + \alpha^{\text{stratum}}_{u[i]} + \alpha^{\text{state}\times\text{stratum}}_{s[i],u[i]} + \alpha^{\text{region}\times\text{stratum}}_{m[i],u[i]})) \\
\alpha^{\text{race}}_{r} &\sim N(0, \sigma^2_{race}), \text{ for } r = 1, 2 \\
\alpha^{\text{region}}_{m} &\sim N(0, \sigma^2_{region}), \text{ for } m = 1, ..., 8 \\
\alpha^{\text{state}}_{s} &\sim N(\alpha^{\text{region}}_{m[s]} + \beta^{\text{RepVote}} * RepVote_s, \sigma^2_{state}), \text{ for } s = 1, ..., 50 \\
\alpha^{\text{stratum}}_{u} &\sim N(0, \sigma^2_{stratum}), \text{ for } u = 1, ..., 7 \\
\alpha^{\text{state}\times\text{stratum}}_{s,u} &\sim N(0, \sigma^2_{state\times stratum}), \text{ for } s = 1, ..., 50 \text{ and } u = 1, ..., 7 \\
\alpha^{\text{region}\times\text{stratum}}_{m,u} &\sim N(0, \sigma^2_{region\times stratum}), \text{ for } m = 1, ..., 8 \text{ and } u = 1, ..., 7 \\
\beta^{\text{RepVote}} &\sim N(0, \sigma^2_{RepVote}), \text{ for } \sigma_{RepVote} = 0.5
\end{aligned}
\tag{2}
$$

[5] On average, in Gallup polls fielded in 1980, 10 states lacked any respondents in each poll, and there is a large, negative correlation between population and unrepresentativeness using observed demographic characteristics. See Appendix B for details.

I drew true effect sizes for the demographic and geographic variables from a normal distribution, varying the standard deviation σ for one variable at a time. For each variable, I ran the simulation with $\sigma \in \{0.1, 0.75, 2.0\}$, holding all other σ values constant at 0.1. As σ increases for each effect, so does the extent to which that variable independently impacts individual opinion. Finally, I use these effects to produce a probability that individual i supports a survey question and draw response y_i from a Bernoulli distribution.

The true data generating process for individual opinion in the simulation is based on race, state, region, and stratum, as well as the interactions between stratum×region and stratum×state, to reflect that that rural and urban places may vary systematically in different parts of the country. $RepVote$ is normally distributed and constrained to be modestly correlated with α_s^{state}.[6]

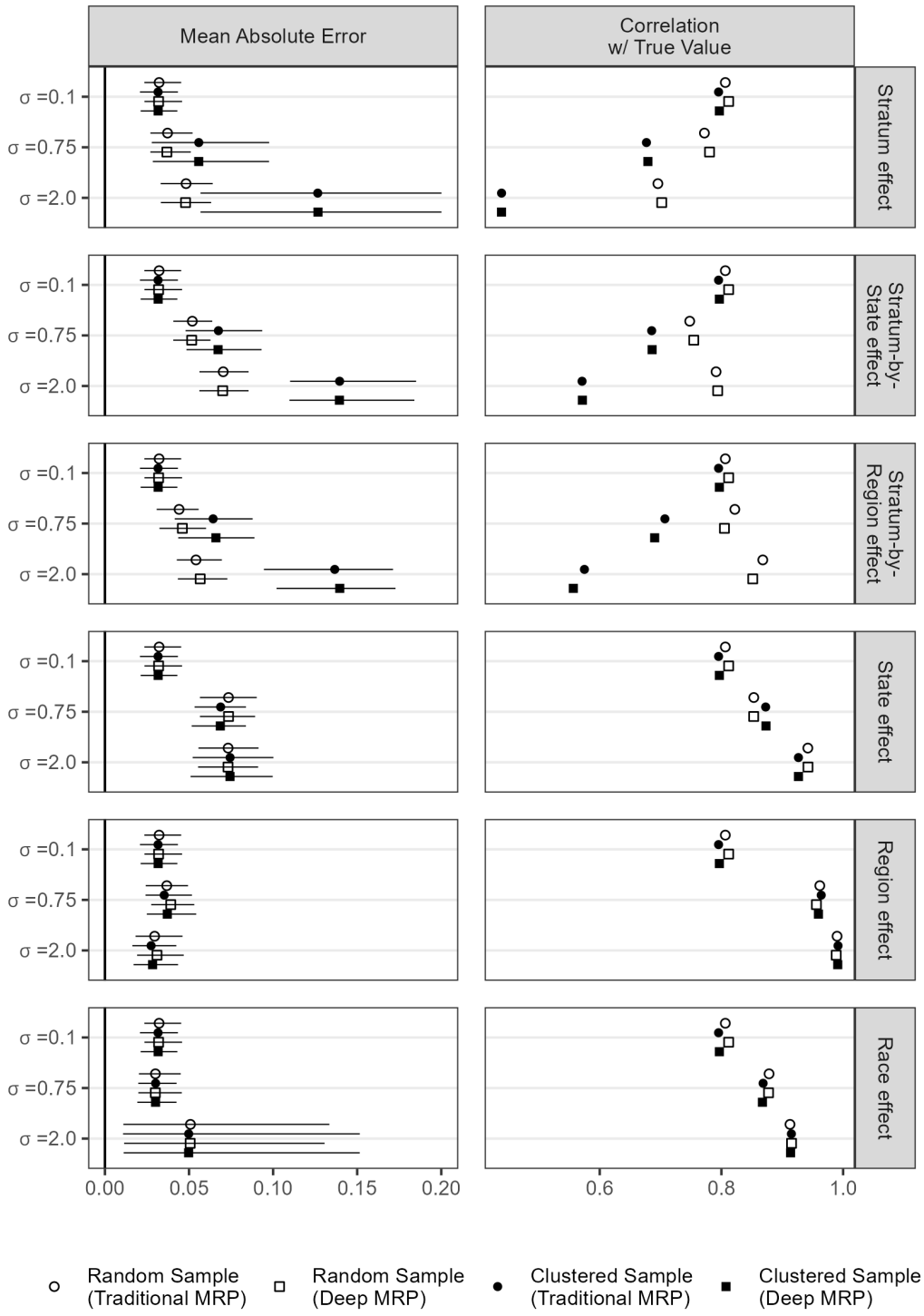
With a population of 1 million voters in hand, I then produce two samples, each meant to mimic a standard survey of approximately 1,500 respondents.[7] The first is a simple random sample in which every voter has an equal probability of being selected. The second is based on the Gallup Poll cluster sampling procedure. For each stratum×region pair, I randomly select two states, weighting by their populations. From each stratum×region×state combination selected, I then sampled 14 respondents from the pool of voters. Finally, I fit both Traditional and Deep MRP models on both sets of polls to predict opinion using the `lme4` package in R (Bates et al. 2015). Deep MRP models add race×state and race×region random effects. I poststratified using all variables and interactions included in the model. I repeated the simulation 100 times for each combination of parameters, allowing the specific effects drawn, the voters, and the samples to vary each time.

Figure 2 reports the results of these simulations. For each variable j, I report error from the MRP model's opinion estimates when the corresponding σ_j is set at 0.1, 0.75, or 2.0. I hold $\sigma_{\neg j}$ for all other variables constant at 0.1. The lefthand column of Figure 2 reports the

[6]The correlation coefficient for $RepVote_s$ and α_s^{state} is on average 0.54. I also replicated simulations with a lower correlation (0.19 on average), and results were similar to those presented.

[7]In fact, my samples have 1,568 respondents each because they have 14 respondents in each of two clusters drawn from each of 56 stratum×region pair.

Figure 2: Results of Simulations: Traditional MRP with Cluster-Sampled Polls



Note: Each point reports results for a series of 100 simulations under a given set of conditions. Simulations vary the standard deviation parameter, σ, for one variable's effect on opinion. All other σ parameters are set to 0.1. Error bars cover results from 95% of simulations.

mean absolute error (MAE) across 100 simulations, using the observed "true" opinion value from the full pool of simulated voters. Filled circles and squares report results from using MRP with the cluster-sampled survey, while the hollow points are for the simple random sample. The righthand column reports the average correlation between between true and modeled opinion.

As I increase the magnitude of the independent effect of stratum on opinion, as well as the the interactions between stratum×state and stratum×region, the amount of error from MRP increases in the clustered random sample. While the MAE does increase slightly for the poll with a simple random sample, the increase in error is much more dramatic when cluster-sampling is used. Notably, MRP does not perform much worse under cluster-sampling when the effects for state, region, or race are large. My results also suggest that although deep models have been found to improve estimation in MRP generally, they do not seem to offer dramatic improvements when polls are cluster-sampled.

We only see divergence between cluster sampling and simple random sampling when the effect of a stratum variable (i.e., heterogeneity inside of subnational units and across the dimensions included in the sampling frame) increases. This suggests that when traditional MRP is conducted with cluster-sampled polls, to the extent that there is heterogeneity inside a state based on stratum, the model performs worse.

2.3 Approaches to MRP with Cluster-Sampled Polls

The abortion case study and simulations above highlight two problems that may produce increased error when estimating state opinion with MRP on cluster-sampled polls.

First, clusters may not be representative of the overall population in a state. At one extreme, some states will have zero respondents despite the fact that a simple random sample might include two or three individuals in expectation. More commonly, they may have a single cluster drawn from just one (unrepresentative) community, as in the Utah case described above. In principle, this is the kind of problem MRP is designed to address (indeed,

even simple random samples will produce states with very few respondents); MRP borrows strength across states and assumes that even if a state has few respondents, a reasonable prediction can be derived from the behavior of similar individuals in other states. However, if clusters are not representative of their states as a whole, this can contaminate the estimate of the state effect and thus lead to inappropriate predictions for the state as a whole.

Second, the hierarchical model typically used with traditional MRP may account for the wrong geographic variation in opinion. The traditional MRP model is inconsistent with the known data generating process for the opinion survey. Because pollsters produced the sample using important information not accounted for in the model, results may have increased error. The traditional MRP model also ignores important information that pollsters use to produce their samples. If a nationally representative survey can be produced in by sampling based on stratum, then these same characteristics may be useful in accurately predicting opinion.

Because of the role that geography plays in a cluster sample, these problems make MRP particularly sensitive to geographic variation in public opinion. Stollwerk (2017) conjectured that MRP estimates from cluster-sampled polls will be incorrect when opinion varies *within* states in ways that are not accounted for by demographic variables. Pollsters produce samples by randomizing not at the state level but at the region×stratum level. As a result, the kinds of people missing from the poll may not always be well represented by those in the dataset. For example, consider the (un)representativeness of an urban neighborhood in Milwaukee being the only cluster sampled in the state of Wisconsin.

In the following sections, I propose and test two approaches to improve estimation of opinion from cluster-sampled surveys. First, I pool responses from multiple surveys. By adding new clusters, the poll in this case begins to approach a simple random sample. (Simple random samples can be thought of as clustered samples with N clusters of 1 respondent each.)

Second, I propose respecifying the predictive model in the first stage of MRP to include the geographic information that pollsters use to produce clustered random samples. This data can then be used in the poststratification step of MRP. Underpinning this approach—

which I call Clustered MRP (CMRP)—is the idea that scholars can improve estimates from MRP by fitting a model that accounts for the cluster-sampling procedure itself.

3 Pooling Cluster-Sampled Polls

One solution to the problem of clustered random samples being unrepresentative of their states as a whole (without conditioning on stratum or cluster-level information) is to pool multiple surveys. MRP performs better with cluster-sampled polls that have a larger number of clusters, which pooling is analogous to, assuming the polls are conducted using different clusters (Stollwerk 2017).[8]

However, pooling surveys improves opinion estimates only in cases where two conditions are met. First, multiple surveys must ask identical (or at least very similar) questions. For many substantive applications (e.g., studying policy responsiveness or constructing time-series of opinion) it may also be necessary for the polls to be fielded around the same time. Second, pooling surveys only produces large reductions in error from MRP when doing so increases the *number* of clusters, and not simply the sample size *within* each cluster. Because survey firms may not change clusters from one poll to the next—and this cannot usually be observed—it is therefore usually necessary to find polls from different firms asking the same questions.

3.1 Simulation Study: Pooling

As an initial test of whether pooling can produce better estimates than using a single cluster-sampled poll, I incorporated pooling into the simulation setup described above. I found that by doubling the number of clusters in each stratum×region combination, estimation can be improved, especially when opinion varies across states or by stratum within states. This is consistent with the problems associated with unrepresentative state-level subsamples from a

[8]See also Pacheco (2013), who follows a pooling approach for multiple surveys during the period that cluster-sampling was common.

single poll. The simulation results are presented and discussed in Appendix D.

However, I also find that pooling does very little to improve opinion estimates if the surveys do not increase the number of clusters. In Appendix D, I show that doubling the sample size within clusters (analogous to pooling two surveys sampling from the same clusters) does not meaningfully improve estimates versus a traditional MRP model.

3.2 Validation: Pooled Surveys for Presidential Opinion

I confirmed the results of the simulation study using presidential election polls from 1980. Presidential elections are a useful testing ground for MRP because they offer a ground truth against which polls can be compared—the election results themselves.

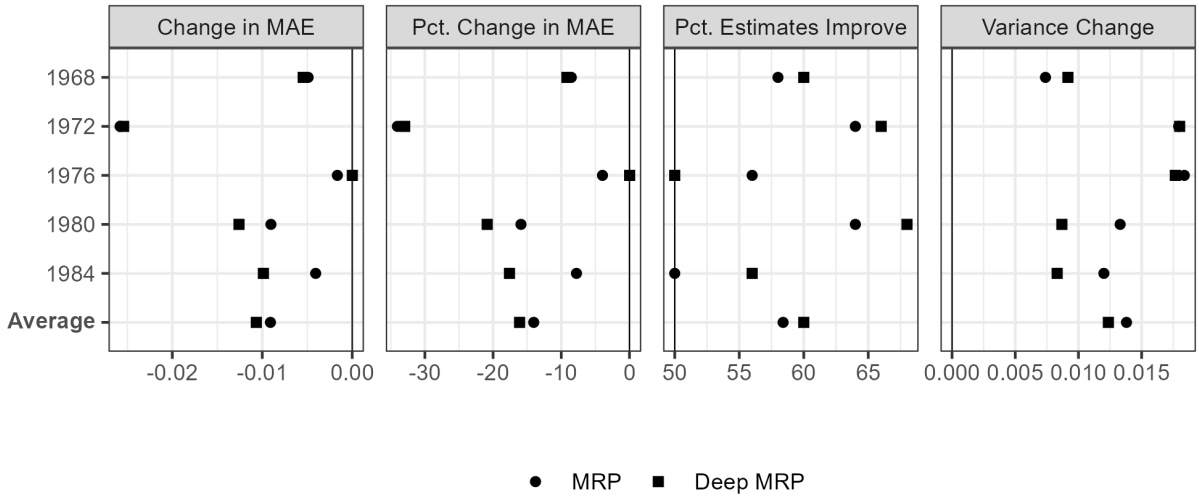
Here, I use MRP to predict support for the Democratic presidential candidate in the 1968–1984 presidential elections. I use two samples: a baseline that comes from the final Gallup poll conducted in the election season, as well as a pooled sample that includes that same Gallup poll and the ANES. For each sample, I fit two models in `Stan`: traditional MRP, which included variables for race, sex, the race×sex interaction, age group, education, and percent evangelical or Mormon;[9] and a Deep MRP model adding interactions among demographic predictors and between demographic and geographic variables. I then poststratified on all included predictors using a poststratification matrix built from joint distributions of the population in the 1980 Census, obtained from IPUMS-NHGIS. The pooled models included a random effect for the survey firm, but I did not poststratify on this variable.[10] The full model specifications and details of the surveys used are in Appendix E.

To account for variability of polls and unobserved changes in the national environment in the final weeks of the campaign (Gelman and King 1993), I report results relative to national support. I adjust both estimates and actual election results by taking the difference between state-level support and national support for the Democratic candidate.

[9]Due to availability of joint population distributions in the Census, the 1968 and 1972 models do not include education.

[10]This approach mirrors that of (Lax and Phillips 2009*a*) who produce a "superpoll" from multiple pollsters

Figure 3: Pooling Presidential Election Polls for MRP



Note: Points represent the improvement observed from using a pooled sample over a single presidential election poll.

Figure 3 reports results. The leftmost column shows the change in MAE that comes from using a pooled sample, rather than the single sample. Negative numbers reflect a reduction in MAE (i.e., more accurate estimation). The second column reports the MAE improvement as a share of the error in the traditional MRP model. The third column shows the share of states whose estimates improved when the pooled sample was used. A pooled sample reduced overall error in four of the five election years—and in the case of 1980 by as much as 30%. Finally, the rightmost panel shows the average increase in the variance of state-level from using a pooled sample (versus standard MRP).[11] The variance of the pooled sample is slightly larger, though not dramatically so, suggesting slightly higher uncertainty from pooling, though the point estimates themselves are more accurate on average.

and include a firm-level random effect.

[11] Average variance is computed by taking the variance for each state's poststratified estimates over 1,000 draws of the posterior distribution of the multilevel model fit in `Stan`. I then average over the variance for all states in each year.

3.3 Limitations of Pooling

While pooling surveys represents a promising improvement on traditional MRP in the case of cluster-sampled polls, there are two challenges that make it impractical in some situations.

First, pooling requires scholars to find multiple polls that asked the same question around the same time. This is often not possible, as many topics appear infrequently in surveys, and the exact question wording can vary widely between polling firms and even from one survey to the next. The need to collect multiple polls can also make scholars' attempts to construct time series of opinion impossible.

Second, simulations indicate that MRP with pooled surveys works well when the number of clusters increases and not necessarily when the size of each cluster does. This makes it especially problematic that polling firms rarely change the communities from which they select respondents, particularly for face-to-face interviews. A review of memos by Gallup statisticians during the 1980s indicated that sampled areas were frequently re-used by pollsters until they had been "exhausted."[12] As a result, the same communities can appear repeatedly in surveys from some pollsters, and because detailed information about the exact location of respondents is generally not available, the extent of this re-use can be difficult to definitively determine. In Appendix D, I show that increasing sample size by doubling the number respondents from each cluster does not offer the same improvements as increasing the number of clusters.

4 Estimating Opinion with Clustered MRP

In this section, I propose Clustered MRP (CMRP), which offers an alternative approach to improving opinion estimation from cluster-sampled polls. CMRP adds geographic data used in the pollster's sampling procedure to both the multilevel model and poststratification stages of MRP. Because clustered random samples are representative of the overall population

[12]These memos were made available by the Roper Center for Public Opinion Research.

conditional on the sampling procedure used, we should be able to improve state opinion estimates by accounting for pollsters' procedures in the model. CMRP also pools respondents more intelligently within regions by allowing missing "types" of people to be represented by more similar groups elsewhere in the sample, rather than dissimilar people in their state. That is, rather than using urban Milwaukee residents to model rural Wisconsinites' opinion, CMRP takes more similar groups (e.g., rural Minnesotans) into account when predicting opinion.

4.1 How to Fit Clustered MRP

CMRP follows a similar procedure to MRP. First, the researcher fits a multilevel model of individual opinion, incorporating the geographic units employed by the pollster to produce the sample. In the Gallup case, this would be region and size-of-community stratum. Specifically, this mirrors the standard MRP approach in Equation (1) above, but adds the below random effects:

$$\begin{aligned}
\alpha_u^{\text{stratum}} &\sim N(0, \sigma^2_{stratum}), \text{ for } u = 1, ..., 7 \\
\alpha_{u,m}^{\text{stratum}\times\text{region}} &\sim N(0, \sigma^2_{stratum\times region}), \text{ for } u = 1, ..., 7 \text{ and } m = 1, ..., 8 \\
\alpha_{u,s}^{\text{stratum}\times\text{state}} &\sim N(0, \sigma^2_{stratum\times state}), \text{ for } u = 1, ..., 7 \text{ and } s = 1, ..., 50
\end{aligned} \tag{3}$$

To improve estimates, we might also include interactions among demographic predictors and between geographic and demographic variables to improve estimation, which I refer to here as Deep CMRP.

Next, the researcher poststratifies to the level of the sampling unit (i.e., stratum) level within each state, using joint distributions from the Census as weights. Finally, the estimates are aggregated up to the state level, again using Census data to weight. The exact steps that need to be taken to produce poststratification information from the Census vary depending on the cluster-sampling procedure used by a polling firm. In general, joint distributions of demographic variables in the population at small geographic levels (e.g., metropolitan areas,

counties, cities and towns) can be downloaded from IPUMS-NHGIS. In some census years, one or two variables may not be included in joint Census tables; however, race, gender, and often age are routinely readily available. The steps I took to produce poststratification matrices used in this paper can be found in Appendix C.

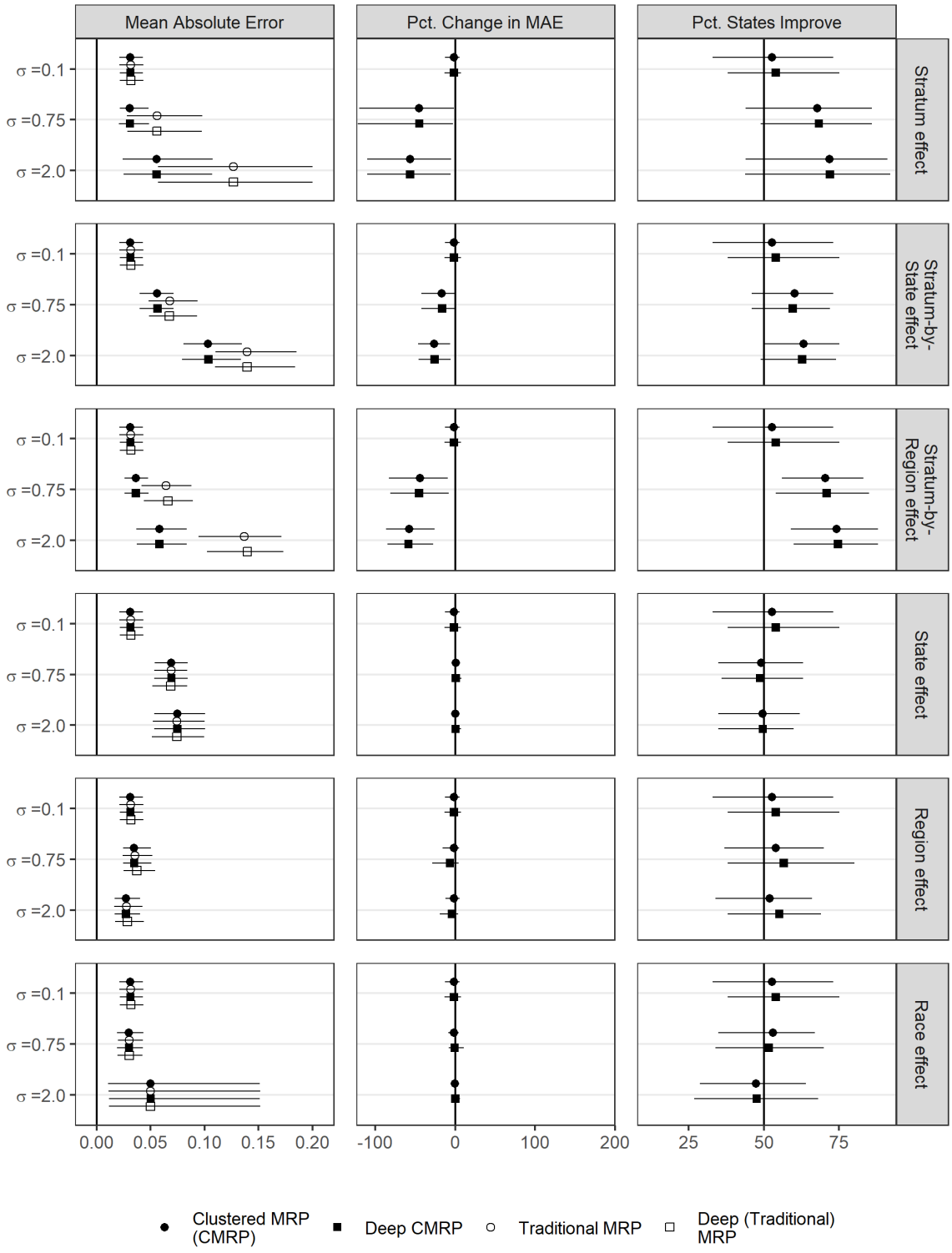
4.2 Simulation Study: Testing CMRP

To test CMRP, I again return to simulations. I follow the same procedure as before but do not generate pooled samples. For each sample, I fit CMRP and Deep CMRP, which adds interactions between race and all geographic variables. I also fit Traditional and Deep MRP models that do not adjust for clustering geography to serve as comparisons. Figure 4 reports the results of these simulations. Here, all results come from polls with clustered random samples. I report results from various CMRP methods (filled circles, triangles, and squares) and the corresponding traditional MRP methods (hollow circles and squares).

The leftmost column of Figure 4 reports the MAE across simulations. As the effect sizes increase for stratum×state, stratum×region, and the independent stratum effect, traditional methods perform worse, while CMRP reduces the error across all specifications. The middle column reports the difference in MAE between corresponding traditional MRP and CMRP methods as a percentage of the error in traditional MRP. A negative result means that the MAE decreases (improves) when CMRP is used. Depending on the conditions shaping opinion, using CMRP might reduce error by as much as 50% when stratum, stratum×state, and stratum×region effect sizes are large. The third column reports the average share of states in each simulation whose modeled estimates of opinion get closer to true opinion. Using CMRP improves the estimates of more than half of states, particularly as the effect sizes for stratum×state, stratum×region, and stratum get larger.

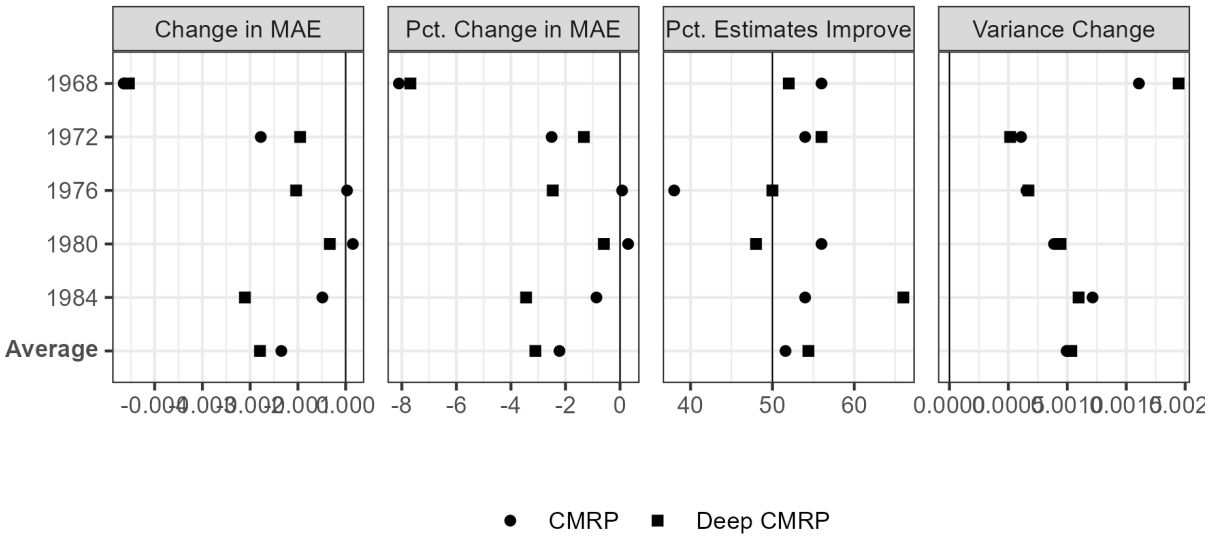
Notably, results differ when the effect of state on opinion is particularly large. As the size of this effect increases, the basic CMRP model without state random effects performs considerably worse than all other models, including Traditional MRP. These simulations

Figure 4: Results of Simulations: Testing Clustered MRP



Note: Points report results for 100 simulations under a set of conditions. Simulations vary the standard deviation σ for one variable's effect on opinion at a time. All models were performed on clustered samples. Percent Change in MAE and Percent of States Improving are relative measures, comparing CMRP to Traditional MRP, and Deep CMRP to Deep MRP. Error bars cover 95% of simulations. Axis limits are constrained to preserve readability.

Figure 5: Validating CMRP with Presidential Election Polls



Note: Points represent the improvement observed from using a given CMRP method on presidential election polls, compared to an analogous traditional MRP method. Traditional MRP is used as a baseline for CMRP; Deep MRP is used as a baseline for Deep CMRP model. Detailed results are in Appendix E.

suggest that CMRP and Deep CMRP should provide the best improvements on traditional MRP when opinion is shaped by geographic variables other than state, while still guarding against poor estimation in cases where state matters a great deal independent of other variables.

4.3 CMRP with Historical Polls

I now turn to validating CMRP using historical cluster-sampled polls. Using CMRP and traditional MRP, I estimated state-level support for Democratic candidates in the five presidential elections from 1968–1984. In each case, I estimate state-level support for the Democratic candidate using the last available Gallup poll before Election Day. I fit two models: CMRP and Deep CMRP, which adds interactions among demographic predictors (e.g., race $\times$ sex $\times$ educ) and between demographic and geographic variables (e.g., race $\times$ state and

race × stratum), which I then compared to similar Traditional MRP or Deep MRP models.[13] As in the pooling example above, all models included variables for race, sex, the race×sex interaction, age group, and percent evangelical or Mormon. Models for 1976–1984 included education (the requisite variables for poststratification were not available in the 1970 Census). I then poststratified on all included predictors using a matrix built from joint distributions of the population in the 1970 and 1980 Censuses, which I obtained from IPUMS-NHGIS. The full model specifications and details of the surveys used are in Appendix E. As before, I report results relative to national support.

Figure 5 reports results. The results indicate that CMRP, on average, reduces the overall error in opinion estimates by 2.5% for the simple CMRP model and 4% for the deep model. These improvements are similar in magnitude to the gains from using machine learning methods in MRP, which have been tested on modern polls that use simple random samples.[14] In some cases, CMRP can produce much larger improvements; using polls from the 1968 election, CMRP performs nearly 10% better than traditional MRP. As predicted in the simulation studies, CMRP and Deep CMRP appear to safeguard against the worst increases in error in the 1972 and 1976 elections. While CMRP outperforms traditional MRP on average, individual state estimates can perform worse in some rare cases, as shown in the third column.

The rightmost column presents differences in variance between estimates produced via CMRP and Traditional MRP. As in the case of pooling, there is slightly more uncertainty associated with the CMRP estimates. However, these differences are much smaller than in the case of pooling—in fact, nearly zero—suggesting that CMRP can improve the accuracy of estimates with limited reduction in precision.

I further tested CMRP for a series of specific issue questions, which are the most common context in which MRP is used but also difficult to validate as high-quality measures of

[13] I compare CMRP with Traditional MRP, and Deep CMRP with a similarly interacted Deep MRP model.

[14] E.g., Ornstein (2020) shows that ensembles improve MRP estimates from standard polls by approximately 2-3%; and Goplerud (2023) finds that Bayesian additive regression trees (BART) outperform standard MRP by 4.5% and deeply specified MRP models by 2.6%.

"ground truth" opinion generally do not exist. In Appendix F, I show that CMRP on average produces slight increases in the correlation between opinion and state liberalism scores by Enns and Koch (2013), though the improvement can be more dramatic on some issues. I also return to the abortion question above and compare it against an abortion liberalism scale, as well as a limited number of state-level public opinion polls in 10 states. I find improvements from using CMRP versus traditional MRP approaches consistent with those reported for estimates of presidential vote choice (around 4% to 10% reductions in MA, depending on the model), though I note that the underlying state polls introduce considerable noise of their own into the comparison. Finally, in Appendix G, I tested CMRP on six issues from the 2000 ANES and compared them with the 2000 National Annenberg Election Survey (NAES). Here, I found that, on average, CMRP methods decrease MAE by 1.2%.

5 Practical Considerations for Clustered MRP

To use CMRP, researchers may need to take additional steps beyond those required for MRP with modern surveys and simple random samples (or similar).

First, in order to model public opinion as a function of the cluster-sampling procedure, it is necessary to obtain more granular individual-level geographic data. Ideally, researchers would fit the model using the exact sampling strata or categories of PSUs from the sampling frame. (Sampling frames and clustering procedures are usually described in the documentation for surveys.) In the case of the Gallup polls that form the core of my validation, the precise stratum designations were not available in the survey data, but a "city size" variable was, which allowed me to match up to Gallup's strata. For some surveys (e.g., the GSS and most years of the ANES), this data is not publicly available and must be requested; in some cases it may not exist at all.

In cases where sufficiently granular data are not available, reasonable proxies may work. In Appendix E, I show that replacing stratum in the Gallup models with a two-level ur-

ban/rural variable produces similar results. I produced this variable using the city size data, as Gallup does not always publish a coarsened urban/rural variable. However, a similar approach may be reasonable for other situations in which samples are produced from clusters based on their urbanness but granular data are not available.

The second major practical consideration is collecting poststratification data necessary for CMRP. Unfortunately, modeling opinion at sub-state levels presents new difficulties not always present in the standard MRP case. While most guides for MRP suggest computing the joint distribution of the population across several demographic variables using Census microdata from IPUMS, the data are too sparse in many smaller geographies to do so.

Instead, joint distributions from tables published by the Census can be used. In this paper, I created poststratification matrices using Census data at the state and place (city or town) level, which I downloaded from IPUMS-NHGIS. For each place, the total population can be used to match to Gallup strata. These can then be combined and aggregated within states to produce the join distribution by race, sex, age, education, and stratum.[15] I discuss this procedure in greater detail in Appendix C. I also have made poststratification data for 1970 and 1980 using the Gallup strata and a simpler urban vs. rural setup in the replication data for this paper.[16]

6 Conclusion

In this paper, I addressed the problem of estimating subnational public opinion using MRP on polls produced from cluster sampling. Simulations suggest that MRP may produce estimates with higher error when clusters are not representative of the overall population of the state and because the multilevel model commonly used in MRP may not correctly account for geographic variation in opinion.

[15]In contexts where joint distributions cannot be contained, it may be possible to estimate them from marginal distributions in the population using multilevel regression with synthetic poststratification (Leemann and Wasserfallen 2017).

[16]Poststratification data can be downloaded from the *State Politics and Policy Quarterly* Dataverse at `URL TO COME`.

To address these potential sources of error, I propose two solutions: pooling samples and CMRP. By pooling multiple cluster-sampled polls that ask the same question around the same time, scholars in effect increase the number of clusters included in the sample frame. By doing so, the MRP model can better account for geographic variation in opinion within and across states. However, this may not be feasible in many contexts.

A second approach—CMRP—improves estimation, on average, without requiring multiple polls. CMRP integrates the sampling procedure used by polling firm into the estimation process by including relevant geographic variables in the predictive model fit in the first step of MRP. Specifically, CMRP fits a multilevel model using demographics and the geographic variables used for clustering (which I call strata, following the Gallup Poll's nomenclature). I also introduced Deep CMRP which adds interactions.

In principle, these two approaches could be combined. That is, one could pool multiple surveys from different firms that use similar sampling procedures, and then produce estimates using CMRP. This poses additional challenges in producing joint distributions of the population for poststratification. However, even in the case where only one method is feasible, the methods in this paper improve estimation of subnational opinion using cluster-sampled polls.

Higher-quality opinion estimates can improve research in a number of domains. First, and most obviously, descriptive studies of public opinion will be aided by more accurate estimates at subnational levels. But public opinion is also useful as an input to understand other political processes and dynamics. Studies of responsiveness depend on estimates of constituent opinion on issues. Likewise, our understanding of party position-taking is often limited by the lack of availability of public support for issues at subnational levels. Reducing measurement error in opinion data may allow for more greater precision in scholarship in in these domains.

In addition to improving estimation from cluster-sampled polls in the U.S. and in comparative contexts, the idea underpinning CMRP may be useful in analyzing opinion data

from other sources with more complex samples. In particular, when selection into a survey varies across some observable other variable, modeling opinion at the level of this variation and aggregating up can reduce measurement error. One concern with online surveys, in particular, has been the unrepresentativeness of samples (Berinsky 2017). MRP has been used to correct unrepresentative online samples in some cases (e.g., Gelman et al. 2016); future research in this vein may be augmented by considering more granular levels at which opinion can be estimated, particularly in cases where observable variables are used deterministically to produce samples.

More generally, the takeaway for scholars is that careful consideration of not only the policy domain at hand but also the procedures used to produce a poll can improve public opinion estimation.

References

Asian Barometer Survey. 2003. "Sampling Procedures of Asian Barometer Survey." http://www.asianbarometer.org/survey/survey-methods.

Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67:1–48.

Berinsky, Adam J. 2017. "Measuring Public Opinion with Surveys." *Annual Review of Political Science* 20(1):309–329.

Brace, Paul, Kellie Sims-Butler, Kevin Arceneaux and Martin Johnson. 2002. "Public Opinion in the American States: New Perspectives Using National Survey Data." *American Journal of Political Science* 46(1):173–189.

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

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. "Stan : A Probabilistic Programming Language." *Journal of Statistical Software* 76(1).

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

Erikson, Robert S., Gerald C. Wright and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. New York: Cambridge University Press.

Gallup Organization. 1980*a*. "Gallup Poll # 1162G, 1980 [Dataset]." https://ropercenter.cornell.edu/ipoll/study/31088016.

Gallup Organization. 1980*b*. "Gallup Poll #1980-1163G: Elections, 1980 [Dataset]." https://ropercenter.cornell.edu/ipoll/study/31088016.

Gelman, Andrew and Gary King. 1993. "Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?" *British Journal of Political Science* 23(4):409–451.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research New York: Cambridge University Press.

Gelman, Andrew, Sharad Goel, Douglas Rivers and David Rothschild. 2016. "The Mythical Swing Voter." *Quarterly Journal of Political Science* 11(1):103–130.

Gelman, Andrew and Thomas C Little. 1997. "Poststratification Into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23(2):127–35.

Ghitza, Yair and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout

and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57(3):762–776.

Goplerud, Max. 2023. "Re-Evaluating Machine Learning for MRP Given the Comparable Performance of (Deep) Hierarchical Models." *American Political Science Review* pp. 1–8.

Grammich, Clifford, Kirk Hadaway, Richard Houseal, Dale E. Jones, Alexei Krindatch, Richie Stanley and Richard H. Taylor. 2019. "Longitudinal Religious Congregations and Membership File, 1980-2010 (State Level)." https://www.thearda.com/Archive/Files/Descriptions/RCMSMGST.asp.

Kish, Leslie and Martin Richard Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society. Series B (Methodological)* 36(1):1–37.

Latin American Pulic Opinion Project. 2019. "AmericasBarometer, 2018/19 Technical Information." https://www.vanderbilt.edu/lapop/ab2018/AmericasBarometer_2018-19_Technical_Report_W_102919.pdf.

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

Lax, Jeffrey R. and Justin H. Phillips. 2009*b*. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science* 53(1):107–121.

Lax, Jeffrey R. and Justin H. Phillips. 2012. "The Democratic Deficit in the States." *American Journal of Political Science* 56(1):148–166.

Leemann, Lucas and Fabio Wasserfallen. 2016. "The Democratic Effect of Direct Democracy." *American Political Science Review* 110(4):750–762.

Leip, David. 2021. "Dave Leip's Atlas of U.S. Presidential Elections.".

Lipps, Jana and Dominik Schraff. 2021. "Estimating Subnational Preferences across the European Union." *Political Science Research and Methods* 9(1):197–205.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler and Steven Ruggles. 2021. "IPUMS National Historical Geographic Information System: Version 16.0.".

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

Pacheco, Julianna. 2013. "The Thermostatic Model of Responsiveness in the American States." *State Politics & Policy Quarterly* 13(3):306–332.

Park, David K., Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–385.

Shirley, Kenneth E. and Andrew Gelman. 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.

Stollwerk, Alissa F. 2017. Essays on the Measurement of Public Opinion Ph.D. diss. Columbia University.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *The Journal of Politics* 75(2):330–342.

Tausanovitch, Chris and Christopher Warshaw. 2014. "Representation in Municipal Government." *American Political Science Review* 108(3):605–641.

Toshkov, Dimiter. 2015. "Exploring the Performance of Multilevel Modeling and Poststratification with Eurobarometer Data." *Political Analysis* 23(3):455–460.

Voss, D. Stephen, Andrew Gelman and Gary King. 1995. "A Review: Preelection Survey Methodology: Details From Eight Polling Organizations, 1988 and 1992." *The Public Opinion Quarterly* 59(1):98–132.

Warshaw, Christopher. 2016. The Application of Big Data in Surveys to the Study of Elections, Public Opinion, and Representation. In *Computational Social Science: Discovery and Prediction*, ed. R. Michael Alvarez. New York: Cambridge University Press pp. 27–50.

Warshaw, Christopher and Jonathan Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *The Journal of Politics* 74(1):203–219.