Bayesian Analysis (2021) **16**, Number 3, pp. 719–744

Improving Multilevel Regression and Poststratification with Structured Priors

Yuxiang Gao*, Lauren Kennedy†, Daniel Simpson‡, and Andrew Gelman§

Abstract. A central theme in the field of survey statistics is estimating population-level quantities through data coming from potentially non-representative samples of the population. Multilevel regression and poststratification (MRP), a model-based approach, is gaining traction against the traditional weighted approach for survey estimates. MRP estimates are susceptible to bias if there is an underlying structure that the methodology does not capture. This work aims to provide a new framework for specifying structured prior distributions that lead to bias reduction in MRP estimates. We use simulation studies to explore the benefit of these prior distributions and demonstrate their efficacy on non-representative US survey data. We show that structured prior distributions offer absolute bias reduction and variance reduction for posterior MRP estimates in a large variety of data regimes.

Keywords: multilevel regression and poststratification, non-representative data, bias reduction, small-area estimation, structured prior distributions, Stan, Integrated Nested Laplace Approximation (INLA).

1 Introduction

Multilevel regression and poststratification (MRP) is an increasingly popular tool for adjusting a non-representative sample to a larger population. In particular, MRP appears to be effective in areas where conventional design-based survey approaches have traditionally struggled, notably small-area estimation (Rao, 2014; Pfeffermann et al., 2013; Zhang et al., 2014) and with convenience sampling (Wang et al., 2015).

One difference between MRP and traditional poststratified design-based weights is that MRP uses partial pooling. Simple poststratification has difficulties with empty cells, in which case the usual practice is to poststratify only on marginals (thus ignoring interactions), or pool cells together. In contrast, the partial pooling of multilevel modeling automatically regularizes group estimates.

Although other options for regularization with poststratification have been explored (Gelman, 2018; Bisbee, 2019), applications of MRP typically assume independent group-level errors. For example, in a political poll modeling there are varying intercepts for states using a regression on region indicators, state-level predictors such as previous

*Department of Statistical Sciences, University of Toronto, Canada, ygao@utstat.toronto.edu
†Columbia Population Research Center and Department of Statistics, Columbia University, New York, NY, lak2183@columbia.edu
‡Department of Statistical Sciences, University of Toronto, Canada, simpson@utstat.toronto.edu
§Department of Statistics and Department of Political Science, Columbia University, New York, NY, gelman@stat.columbia.edu

 https://doi.org/10.1214/20-BA1223

voting patterns in the state, plus independent errors at the state level. In some applications though, there is potential benefit from including underlying structure not captured by regression predictors. We demonstrate that this structure can be captured through more complex prior specifications. For example, instead of independent errors for an ordered categorical predictor, we specify an autoregressive structure instead. Ordered predictors are just one example where we can introduce *structured prior distributions*.

We begin this manuscript by providing an overview of the existing literature on methods that adjust for nonrepresentative data, as well as a review on the common application areas of MRP.

1.1 Existing Literature on Post-Sampling Adjustments for Non-Representativeness and MRP

Post-sampling adjustments aim to correct for differences between a potentially biased sample and a target population. Poststratification is a commonly used weighting procedure for nonresponse in model-based survey estimates (Little, 1993). It can improve accuracy of estimates but is no silver bullet, since the quality of poststratified estimates depends on the quality of the known information about the population sizes of the strata, along with the assumption that the sample is representative of the population within each poststratification cell. An approximation to poststratification is raking, which is an iterative algorithm using marginal totals (Deming and Stephan, 1940; Lohr, 2009; Skinner et al., 2017). When adjusting for many factors, raking can yield unstable estimates caused by high variability of the adjusted weights (Izrael et al., 2009). For a modern overview of current methods of inference and post-sampling adjustments for nonprobability samples, see Elliott et al. (2017). As the demands for small-area estimation increase, so too should the utility of MRP. We use structured priors in our proposed improvement for MRP, with the aim of more sensible shrinkage of posterior estimates that should ultimately reduce estimation bias.

MRP has been used in a broad range of applied problems ranging from epidemiology (Zhang et al., 2014; Downes et al., 2018) to social science (Lax and Phillips, 2009; Wang et al., 2015; Trangucci et al., 2018). MRP's beginnings saw applications in political science (Gelman and Little, 1997; Park et al., 2004) for the estimation of state-level opinions from national polls. The breadth of its applications has since matured substantially, even to the extent of being used by data journalists (Morris, 2019). One of MRP's appeals to applied researchers is the ability to produce reliable estimates for small areas in the population and simultaneously adjust for non-representativeness.

On the methodology front, Ghitza and Gelman (2013) and Gelman et al. (2016) extended MRP to include varying intercepts and slopes for interactions, along with inference for time series of polls.

1.2 Outline for This Paper

This work explores alternative regularization techniques with structured prior distributions that lead to absolute bias reduction in MRP estimates. Our methodology of

structured priors should not be confused with that of Si et al. (2017), who define structured priors as a way to perform variable selection for higher-order interaction terms. Our improvements on estimation precision come from replacing independent distributions of varying coefficients with Gaussian Markov random fields (Rue and Held, 2005).

This paper is structured as follows: Section 1 provides an overview of the existing literature for MRP and nonrepresentative data. Section 2 gives a concise overview of MRP and what's required for the methodology. Section 3 describes our structured priors framework in detail, along with motivation for their use in MRP. Section 4 presents simulation studies of structured priors across various regimes of non-representative survey data. Explanation of the simulation setup and interpretation of the simulation results are given. Bias and variance comparisons are made between structured priors and the classical independent random effects in MRP in section 4. Section 5 contains the application of structured priors in MRP to a real survey data set that's non-representative. Section 6 is the conclusion. All computation carried out in this manuscript was done in R (R Core Team, 2019).

2 Overview of MRP

Multilevel regression and poststratification Gelman and Little (1997) proceeds by fitting a hierarchical regression model to survey data, and then using the population size of each poststratification cell to construct weighted survey estimates. More formally, suppose that the population contains K categorical variables and that the k^{th} has J_k categories. Hence the population can be represented by $J = \prod_{k=1}^{K} J_k$ cells. Usually the population contains continuous variables, and in that case these variables will be discretized to form categorical variables. For example, age in a demographic study can be discretized into a finite number of categories. For every cell, there is a known population size N_j. Increasing the number of groups for a continuous variable will increase the number of cells J and correspondingly decrease the individual cell population sizes N_j.

Choosing the optimal group size for continuous variables is a difficult model selection problem, involving tradeoffs between accuracy and computational load, and this is something that we do not address in this manuscript. This area of MRP research is an active field and the best practice has yet to emerge from the research.

Suppose that the response variable of individual i is $y_i \in \{0, 1\}$. MRP for binary survey responses is summarized by the two steps below:

Multilevel regression step Let n be the number of individual observations in a dataset. Fit the hierarchical logistic regression model below to get estimated population averages θ_j for every cell $j \in \{1, \ldots, J\}$. The hierarchical logistic regression portion of MRP has a set of varying intercepts $\{\alpha_{j*}^{k}\}_{j*=1}^{J_k}$ for each categorical covariate k, which have the effect of partially pooling each θ_j towards a globally-fitted regression model, $X_j\beta$, with sparse cells benefiting the most from this regularization. X_j is the row in the design matrix that corresponds to θ_j, where $j \in \{1, \ldots, J\}$. We follow a notation consistent

with Gelman and Hill (2006).

$$\begin{aligned}
\Pr(y_i = 1) &= \text{logit}^{-1}\left(X_i\beta + \sum_{k=1}^{K} \alpha^k_{j[i]}\right), \text{ for } i = 1, \ldots, n, \\
\alpha^k_j \mid \sigma^k &\overset{\text{ind.}}{\sim} \text{N}(0, (\sigma^k)^2), \text{ for } k = 1, \ldots, K, \ j = 1, \ldots, J_k, \\
\sigma^k &\sim \text{N}_+(0, 1), \text{ for } \text{ for } k = 1, \ldots, K, \\
\beta &\sim \text{N}(0, 1),
\end{aligned}$$

where we are giving default weakly informative priors to the non-varying regression coefficients β.

Poststratification step Using the known population sizes N_j of each cell j, poststratify to get posterior preference probabilities at the subpopulation level. The poststratification portion of MRP adjusts for nonresponse in the population by taking into account the sizes of every cell l relative to the total population size $N = \sum_{j=1}^{J} N_j$. Another way to interpret poststratification is as a weighted average of cell-wise posterior preferences, where the weighting scheme is determined by the size of each cell in the population. Smaller cells get downweighted and larger cells get upweighted. The final result is a more accurate estimate in the presence of non-representative data.

Let S be some subset of the population defined based on the poststratification matrix. Then the poststratified estimand for S is:

$$\theta_S := \frac{\sum_{j \in S} N_j \theta_j}{\sum_{j \in S} N_j}.$$

For example, S could correspond to the oldest age category in the lowest income bracket. Then θ_S would correspond to the proportion of people in this sub-population that would respond yes to the survey question of interest. It's important to note that, one can model at a finer scale than the poststratification scale.

3 Proposed Approach and Motivation

We consider structured prior distributions for MRP taking the form of Gaussian Markov random fields (GMRF), modeling certain structure of the underlying categorical covariate in the hierarchical regression. We proceed as follows for a covariate in the population of interest:

Case 1. If we do not want to model any structure in a categorical covariate, we model its varying intercepts as independently normally distributed. This would be what's described in the previous section, Overview of MRP.

Case 2. If there is underlying structure we would like to model in a covariate, and spatial smoothing using this structure seems sensible for the outcome of interest, then we use an appropriate GMRF as a prior distribution for this batch of varying intercepts.

We will specify informative hyperpriors when possible and model via a full Bayesian approach. For a detailed overview of principled hyperprior specification in GMRF models, we refer the reader to Simpson et al. (2017). As well, we do not restrict structured priors to have directed or undirected conditional distributions (Rue and Held, 2005). Some examples of directed conditional distributions include the autoregressive and random walk processes with discrete time indices, which are frequently used in time series analysis. The Conditional Autoregressive (CAR) and Intrinsic Conditional Autoregressive (ICAR) processes (Besag, 1975) are common undirected conditional distributions and are often used in specifying priors in spatial models.

More complex prior structure allows for nonuniform information-borrowing in the presence of non-representative surveys from a population. For example, it makes sense to partially pool inferences for the oldest age group toward data from the second-oldest group. An autoregressive prior placed on the ordinal variable age achieves this effect, without making the strong global assumptions involved in simply including age as a linear or quadratic predictor in the regression. The proposal of using structured priors aims to reduce bias for MRP estimates in extremely non-representative data regimes.

Structured priors improve upon the multilevel aspect of MRP while maintaining the regression structure. Because MRP is a model-based survey estimation approach, the multilevel regression component can be replaced with other forms of regression modelling, for example with sparse hierarchical regression (Goplerud et al., 2018) or Bayesian additive regression trees (Bisbee, 2019). It is important, though, that the regression step be regularized in some way to preserve the ability of the method to account for a potentially large number of adjustment factors and their interactions (Gelman, 2018). When compared to machine learning-style regularization methods as seen in Bisbee (2019) and Goplerud et al. (2018), structured priors offer a lot more interpretability in the modelling step of MRP. For example, it's not as clear how Bayesian additive regression trees (BART) allow for information-borrowing for structured covariates despite BART and related regularized tree-based methods (Chen and Guestrin, 2016) being modern methods in out-of-sample prediction. In contrast to this, an AR(1) prior with autoregression coefficient $\rho \in [0, 1)$ on the ordinal variable age has the clear interpretation that posterior estimates are regularized towards their previous first-order neighbouring age, where the amount of regularization is determined by the coefficient ρ.

GMRFs have a deep connection with Gaussian Processes (GPs). As the discretization of the underlying space gets finer, under certain technical conditions, a GMRF will converge to a specific GP (Lindgren et al., 2011). More details on the style of convergence can be found in Lindgren et al. (2011). These theoretical foundations show that structured priors in MRP are a discrete approximation to GP priors for structured covariates. GPs are universal function approximators, hence structured priors in MRP can be thought of as a flexible modeling strategy that simultaneously takes into account various data-generating processes and offers interpretable regularization.

4 Simulation Studies

4.1 Directed Structured Priors Example: Models for Partial Pooling of Group-Level Errors

For the first simulation example, we work with a simple model of three poststratification categories—51 states, age in years ranging from 21–80, and income in 4 categories—and no other predictors. Age is further categorized into 12 groups. In practice, the number of categories that age is discretized into for MRP models is a lot smaller than the total number of possible integer ages (Lax and Phillips, 2009; Trangucci et al., 2018). 12 categories were chosen since it was large enough to allow for prior distribution structure to introduce intelligent information-borrowing in the age covariate yet it was a lot less than the maximal number of age categories, 60. We define $\alpha^{\text{Age Cat.}}_{j[i]}$, $\alpha^{\text{Income}}_{j[i]}$ and $\alpha^{\text{Region}}_{j[i]}$ to be the varying intercepts for age category, income category and region respectively for the i^{th} survey respondent. Phone surveys will often have these three covariates on respondents, often times with the nonrepresentativeness of the survey driven by varying nonresponse rates across ages in the population.

For all three prior specifications of MRP, we use the link function,

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \alpha^{\text{State}}_{j[i]} + \alpha^{\text{Age Cat.}}_{j[i]} + \alpha^{\text{Income}}_{j[i]}\right), \text{ for } i = 1, \ldots, n. \tag{4.1}$$

For all three prior specifications we assume independent mean-zero normal distributions for the $\alpha^{\text{Region}}_{j[i]}$'s, $\alpha^{\text{Age Cat.}}_{j[i]}$'s and $\alpha^{\text{Income}}_{j[i]}$'s along with a weakly informative half-normal distribution for the corresponding scale parameter:

$$\begin{aligned}
\alpha^{\text{State}}_j \mid \beta^{\text{State-VS}}, \beta^{\text{Relig.}}, (\alpha^{\text{Region}}_{j^*})^5_{j^*=1}, \sigma^{\text{State}} &\overset{\text{ind.}}{\sim} \text{N}(\alpha^{\text{Region}}_{m[j]} + \beta^{\text{Relig.}} X_{\text{Relig.},j} \\
&\quad + \beta^{\text{State-VS}} X_{\text{State-VS},j}, (\sigma^{\text{State}})^2), \\
&\quad \text{for } j = 1, \ldots, 51, \\
\alpha^{\text{Region}}_m \mid \sigma^{\text{Region}} &\overset{\text{ind.}}{\sim} \text{N}(0, (\sigma^{\text{Region}})^2), \\
&\quad \text{for } m = 1, \ldots, 5, \\
\alpha^{\text{Income}}_j \mid \sigma^{\text{Income}} &\overset{\text{ind.}}{\sim} \text{N}(0, (\sigma^{\text{Income}})^2), \\
&\quad \text{for } j = 1, \ldots, 4, \\
\sigma^{\text{Income}}, \sigma^{\text{State}}, \sigma^{\text{Region}} &\sim \text{N}_+(0, 1), \\
\beta^{\text{State-VS}}, \beta^{\text{Relig.}}, \beta^0 &\sim \text{N}(0, 1),
\end{aligned} \tag{4.2}$$

where $X_{\text{State-VS},j} \in [0, 1]$ is the covariate that corresponds to the 2004 Democratic vote share for state j and $X_{\text{Relig.},j} \in [0, 1]$ is the percentage of conservative religion in state j, which is defined as the sum of the percentage of Mormons and percentage of Evangelicals in state j. The term $\alpha^{\text{Region}}_{m[j]} + \beta^{\text{Relig.}} X_{\text{Relig.},j} + \beta^{\text{State-VS}} X_{\text{State-VS},j}$ are state-level predictors that utilize auxillary data accounting for structured differences among the states.

The *baseline specification* is the classical prior distribution used in MRP with independent normal distributions for the varying intercepts for age categories:

$$\begin{aligned} \alpha_j^{\text{Age Cat.}} \mid \sigma^{\text{Age Cat.}} &\overset{\text{ind.}}{\sim} \text{N}(0, (\sigma^{\text{Age Cat.}})^2), \text{ for } j = 1, \ldots, 12, \\ \sigma^{\text{Age Cat.}} &\sim \text{N}_+(0, 1). \end{aligned} \tag{4.3}$$

The *autoregressive specification* models the ordinal structure of age category as a first-order autoregression (Rue and Held, 2005). The prior distribution imposed on ρ is restricted to the range $(-1, 1)$, enforcing stationary for the autoregressive process.

$$\begin{aligned} \alpha_1^{\text{Age Cat.}} \mid \rho, \sigma^{\text{Age Cat.}} &\sim \text{N}(0, \frac{1}{1-\rho^2}(\sigma^{\text{Age Cat.}})^2), \\ \alpha_j^{\text{Age Cat.}} \mid \alpha_{j-1}^{\text{Age Cat.}}, \ldots, \alpha_1^{\text{Age Cat.}}, \rho, \sigma^{\text{Age Cat.}} &\sim \text{N}(\rho\alpha_{j-1}^{\text{Age Cat.}}, (\sigma^{\text{Age Cat.}})^2), \\ &\quad \text{for } j = 2, \ldots, 12, \\ \sigma^{\text{Age Cat.}} &\sim \text{N}_+(0, 1), \\ (\rho + 1)/2 &\sim \text{Beta}(0.5, 0.5). \end{aligned} \tag{4.4}$$

Finally, we consider the *random walk specification*, which is a special case of first-order autoregression with ρ fixed as 1, although with a different parameterization to avoid the division by $1-\rho^2$ above. In addition, we introduce the sum-to-zero constraint $\sum_{j=1}^{J} \alpha_j^{\text{Age Cat.}} = 0$ to ensure that the joint distribution for the first-order random walk process is identifiable.

$$\begin{aligned} \alpha_j^{\text{Age Cat.}} \mid \alpha_{j-1}^{\text{Age Cat.}}, \ldots, \alpha_1^{\text{Age Cat.}}, \sigma^{\text{Age Cat.}} &\sim \text{N}(\alpha_{j-1}^{\text{Age Cat.}}, (\sigma^{\text{Age Cat.}})^2), \\ &\quad \text{for } j = 2, \ldots, 12, \\ \sigma^{\text{Age Cat.}} &\sim \text{N}_+(0, 1), \\ \sum_{j=1}^{J} \alpha_j^{\text{Age Cat.}} &= 0. \end{aligned} \tag{4.5}$$

The three prior specifications differ in the amount of information shared between neighbors in the age category random effect. In the baseline specification, no information is shared between $\alpha_j^{\text{Age Cat.}}$ and $\alpha_{j-1}^{\text{Age Cat.}}$ for $j \in \{2, \ldots, 12\}$. In the autoregressive specification, partial information is shared and in the random walk specification the full amount of information is shared. The sharing of information between $\alpha_j^{\text{Age Cat.}}$ and $\alpha_{j-1}^{\text{Age Cat.}}$ is analogous to shrinkage of one posterior towards another. In this case, the posterior of $\alpha_j^{\text{Age Cat.}}$ shrinks toward the posterior of $\alpha_{j-1}^{\text{Age Cat.}}$ under both the random walk and autoregressive specifications. The amount of shrinkage is governed by the autoregressive coefficient ρ. The reason behind specifying first-order autoregressive processes as the structured prior for age is that individuals with similar ages should have similar opinions for the survey question of interest. First-order autoregressive processes incorporate the prior assumption that the opinion of an individual for a certain age is similar to the opinion of individuals with exactly the same demographics except with a slightly younger age. In the simulation studies below, we empirically show that the

property of shrinking towards the previous neighboring variable in the autoregressive and random walk specifications result in decreased posterior bias of MRP estimates for every cell in the population.

Simulated Data

The sample We consider three scenarios of true $\mathrm{E}(y)$ as a function of age: U-shaped, cap-shaped, or monotonically increasing. We investigate the effects of non-representative data amongst elderly individuals (ages 61–80) in the simulation samples, and show that the random walk specification provides the lowest absolute bias in subpopulation level estimates when compared to the other two specifications. The likelihood of sampling from a subpopulation group given that individuals respond is dependent on the size of the subpopulation group along with the response probability of an individual in that group.

The probability vector of sampling is defined as:

$$\frac{(\text{Probability of response}) \odot (N_1, \ldots, N_J)}{\sum_j (\text{Probability of response})_j \cdot N_j},$$

where $\odot$ is the Hadamard product. This probability vector is in reference to the post-stratification matrix defined for this simulation study. A special case for the probability vector of sampling is when the probability of response is equal for all cells in the population, resulting in a probability vector of sampling that's fully representative of the population. The probability vector of sampling is used to generate a sample of binary responses along with covariates. Through this probability vector, one can augment it to get highly non-representative samples for certain subpopulation groups. In the case of a completely random sample for subpopulation groups of interest, all subpopulation groups of interest have the same probability of sampling. As an example, all 12 age categories would have equal probability of being sampled from in the scenario of completely random sampling for age categories.

Assumed sample and population In the following simulation study we will assume that the population is sufficiently large so that sampling with replacement is equivalent to retrieving a random sample from the population.

To empirically validate the improvements that structured priors have on posterior MRP estimates, we construct various data regimes for age categories 9–12. More specifically, let S be the index set corresponding to age categories 9–12. Summing the probability of sampling over S will return the expected proportion of the sample who are older adults. We perturb this probability through 9 scenarios, ranging from 0.05 (under-representing older adults) to 0.82 (over-representing older adults). This section contains plots for the U-shaped true preference curve, with the appendix containing plots for the increasing-shaped true preference curve and the cap-shaped true preference curve.

These three true preference curves capture the rough structure of the unseen truths in real survey data. Let x represent age of an individual, and let $f(x)$ represent the preference curve for age. The three different preference curves with respect to age are

defined as:

Cap-shaped preference:

$$f(x) = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} x(1-x),\ x \in [21, 80].$$

U-shaped preference:

$$f(x) = 1 - \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} x(1-x),\ x \in [21, 80].$$

Increasing-shaped preference:

$$0.7 - 3\exp(-\frac{x}{0.2}),\ x \in [21, 80].$$

True preferences for every poststratification cell $j \in \{1, \ldots, J\}$ in the population are then generated with the following formula:

$$\theta_j = \text{logit}^{-1}\big(\beta^0 + f(X_{\text{Age}[j]}) + X_{\text{Income}[j]} + \beta^{\text{State}} X_{\text{State}[j]} + \beta^{\text{Relig.}} X_{\text{Relig.}[j]}\big),$$

where $X_{\text{Age}[j]}$, $X_{\text{Income}[j]}$, $X_{\text{State}[j]}$, $X_{\text{Relig.}[j]}$ correspond to the age, income effect, state effect and religion effect respectively of poststratification cell j. $X_{\text{Income}[j]}$, $X_{\text{State}[j]}$, $X_{\text{Relig.}[j]}$ along with $\beta^0, \beta^{\text{State}}$ and $\beta^{\text{Relig.}}$ are defined in the appendix (Gao et al., 2020).

Directed Structured Priors Results

We fit all models using the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2016) to perform full Bayesian inference, using the default settings of 2000 iterations on 4 chains run in parallel, with half the iterations in each chain used for warmup.

Impact of prior choice on bias of posterior preferences The first way we evaluate the impact of prior specification is by considering the impact of bias when we manipulate the expected proportion of the sample that are older adults. In Figure 1 below, we plot the results for a sample size of 100 and 500.

When the expected proportion of the sample that are older adults is equal to 0.33, this corresponds to a completely random sample for age categories (probability of sampling every age category is the same) *and* a fully representative sample for age categories (probability of sampling every age category is proportional to the population sizes for every age category). In certain scenarios, a completely random sample may be more desirable than a fully representative sample of the population for modeling purposes. Certainly, oversampling a sparse subpopulation group in the population will return lower variance model estimates for that specific subpopulation group.

We can see from Figure 1 that the two structured prior specifications outperform the baseline prior specification by a few percentage points for almost all 12 age categories, and achieving the same performance for the remaining age categories.

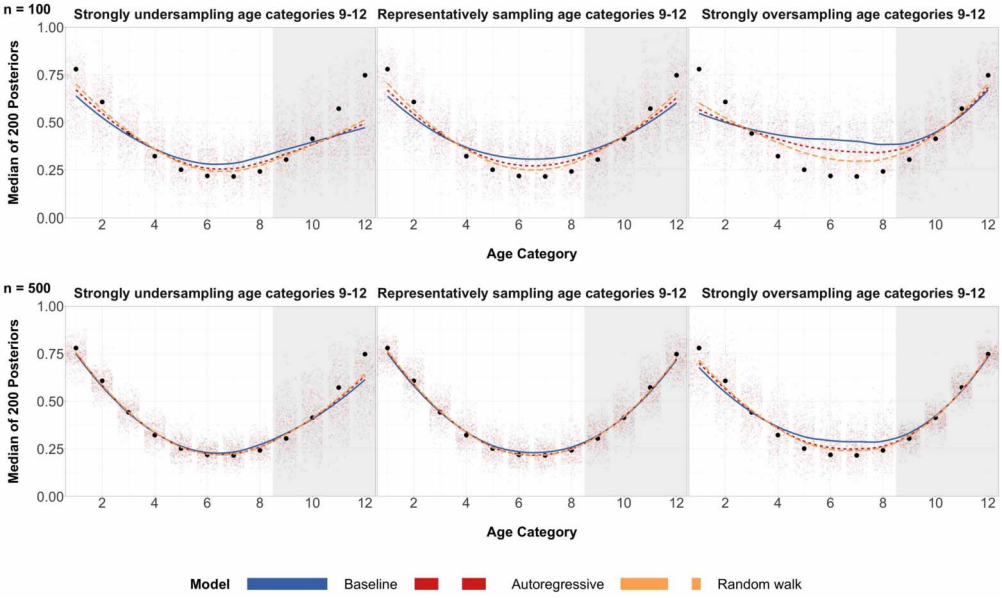

Figure 1: Posterior medians for 200 simulations for each age group under three different regimes of data, where true age preference is U-shaped. The top row corresponds to a sample size of 100 and the bottom row corresponds to a sample size of 500. Black circles are true preferences for each age group. The shaded grey region corresponds to the age categories of older individuals for which we over/undersample. The left column has a probability of sampling age categories 9–12 equal to 0.05. The middle column has a probability of sampling age categories 9–12 equal to 0.33, which is completely random sampling *and* representative sampling for all age categories. The right column has a probability of sampling age categories 9–12 equal to 0.82. Local regression is used for the smoothed estimates amongst the three prior specifications. For the same plots involving different probabilities of sampling, refer to Appendix Table 1.

When elderly individuals are undersampled relative to the rest of the population, the random walk prior specification outperforms the baseline prior specification in lower absolute bias by a few percentage points across all the age categories.

When elderly individuals are oversampled relative to the rest of the population, the random walk prior specification outperforms the baseline prior specification in lower absolute bias by close to 10 percentage points for mid-aged individuals when sample size is 100. As expected, the three prior specifications produce essentially the same posterior estimates in the bottom row of Figure 1, due to the sample size being large in each of these age categories – Increasing n will increase the weight of the likelihood on the posterior in a statistical model. Regardless, absolute bias is reduced or stays the same for all age categories and all data regimes for the two structured priors specifications, as seen in Figure 1.

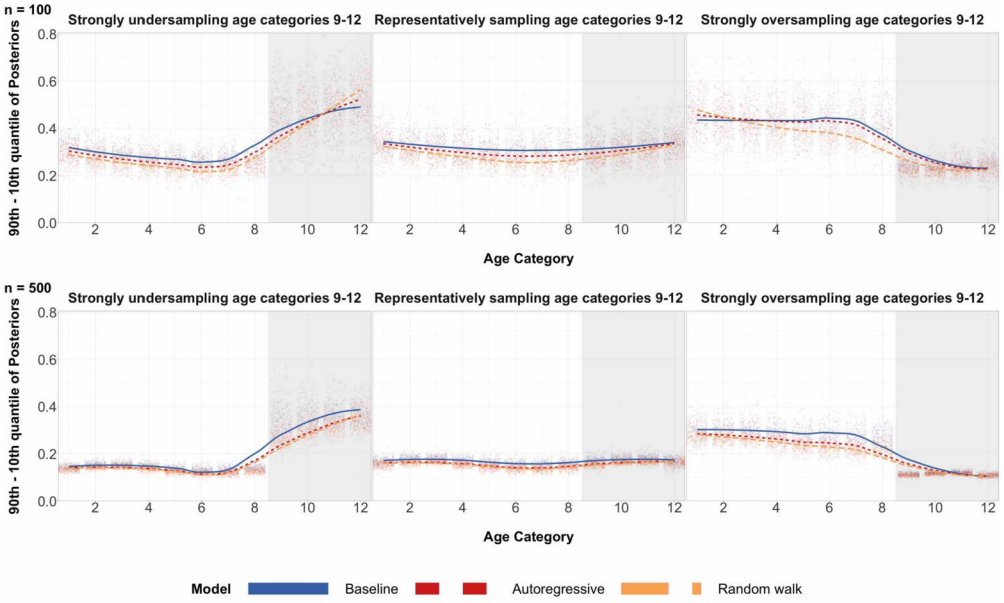

Figure 2: Differences in the 90^{th} and 10^{th} posterior quantiles for every age category when true preference is U-shaped for 200 simulations. The top row corresponds to a sample size of 100 and the bottom row corresponds to a sample size of 500. The shaded grey region corresponds to the age categories of older individuals for which we over/undersample. The left column has a probability of sampling age categories 9–12 equal to 0.05. The middle column has a probability of sampling age categories 9–12 equal to 0.33, which is completely random sampling *and* representative sampling for all age categories. The right column has a probability of sampling age categories 9–12 equal to 0.82. Local regression is used for the smoothed estimates amongst the three prior specifications. For the same plots involving different probabilities of sampling, refer to Appendix Table 2.

As a secondary benefit of structured priors, averaging over all 200 runs, simulation studies had shown the difference of the 90^{th} and 10^{th} posterior quantiles for almost all age categories to be smaller when $n = 100$. This is shown in Figure 2. This difference can be interpreted as a measure of posterior standard deviation. When $n = 500$, reduction in posterior quantiles difference is even more apparent. Reduction in posterior standard deviations may not be ideal for estimators when the tradeoff is higher absolute bias, but for the case of structured priors, we see a reduction in both for every age category implying a decrease in L_2 risk for posterior estimates of every age category.

Another visualization of bias reduction is based on Figure 3. It shows bias of posterior preferences for each cell in the population, the finest granularity, as the expected proportion of the sample that are older adults is perturbed. Absolute bias is significantly decreased when switching from the baseline specification to the random walk specifi-

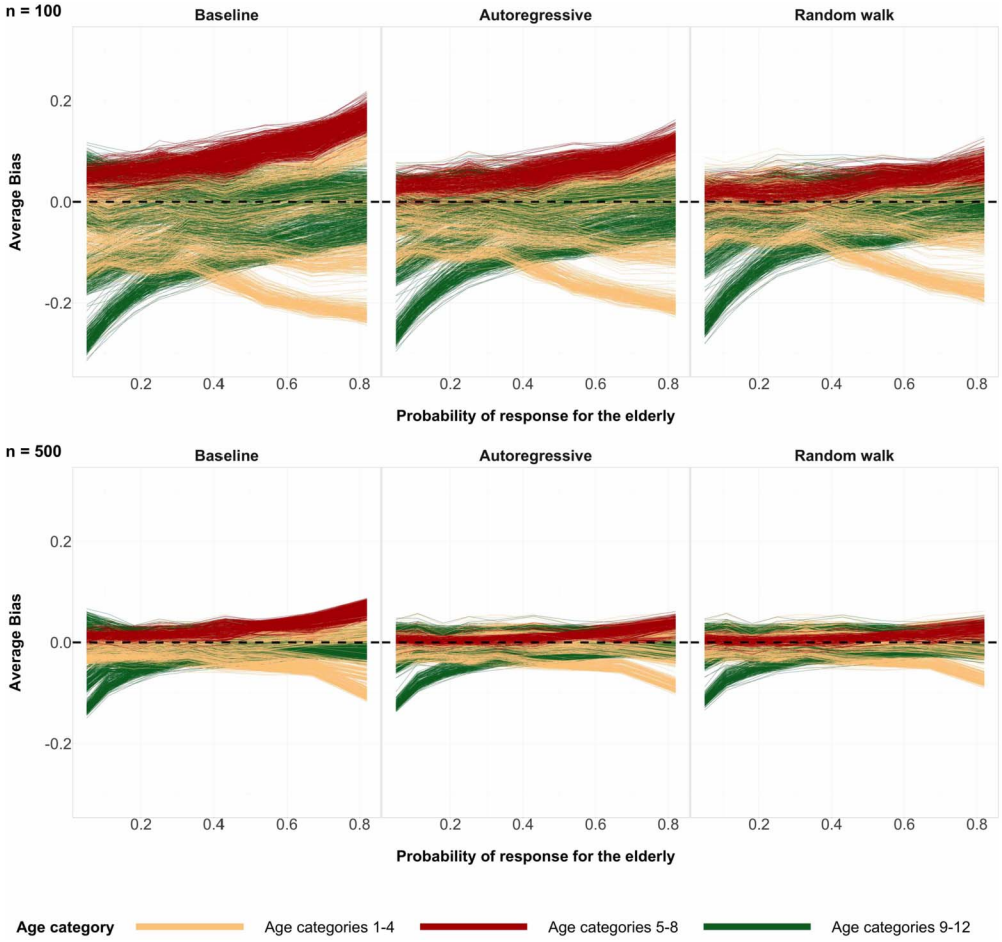

Figure 3: The average bias values coming from 200 simulations of posterior medians of the 2448 poststratification cells. The possible values of average bias are in the interval $(-1, 1)$. Sample size $n = 100$ (top) and $n = 500$ (bottom). The true preference curve for age is U-shaped. The horizontal dashed line at $y = 0$ represents zero bias.

cation. The autoregressive specification also reduces absolute bias, but not as much when compared to the random walk specification. This is due to the prior ρ defining before inference that the information being borrowed from the neighboring age category posteriors should be a value in $[-1, 1]$.

The population preference estimates for the three prior specifications remain nearly the same across all probability of sampling indices when the true preference curve is U-shaped or cap-shaped. When the true preference curve is increasing-shaped, the population preference remains nearly the same for all probability of sampling indices except 0.05 and 0.82. In those cases, the first-order random walk prior produces more unbiased

population estimates by a few percentage points. The advantage of structured priors appears to be more drastic when reducing to more granular sub-population levels. For additional bias plots on all three true preference curves, the reader can refer to the appendix.

In summary, based on the simulation studies on the U-shaped true preference along with the two other true preference curves, we see that structured priors decrease absolute bias for posterior MRP estimates more than the classical specification of priors in MRP, regardless of how representative the survey data are to the population of interest. This implies that posterior MRP estimates coming from structured priors are much more invariant to differential nonresponse and biased sampling when compared to the classical priors used in MRP. The main goal of this paper is to argue that structured priors offer an improvement to MRP, even in extremely non-representative data regimes. We indeed see that in the simulation studies as large decreases in absolute bias are seen when the probability of sampling age categories 9–12 are 0.05 and 0.82. A secondary benefit of structured priors is variance reduction on posterior estimates of the structured covariates.

The benefits of structured priors in this case are apparent even when sample size is 1000, but to a lesser degree than when sample size is 100 or 500. The absolute bias reduction and posterior standard deviation reduction due to structured priors when $n = 1000$ are a few percentage points for the majority of under/over-represented age categories, based on all three true preferences. In contrast to this, the absolute bias reduction and posterior standard deviation reduction due to structured priors when $n = 100$ are around 10 percent for the majority of under/over-represented age categories. For results on this $n = 1000$ simulation, refer to Appendix Table 1 and Appendix Table 2 with the configuration $n = 1000$ for all three true preference shapes.

The three prior specifications start to converge in terms of having the same posterior bias and variance with an increase in sample size due to the likelihood dominating the prior distributions imposed. Regardless, with a sample size of 1000 and 12 age categories, it appears that structured priors are still able to reduce posterior variance and absolute bias in the nonrepresentative data regime – where elderly individuals are oversampled and undersampled.

The efficacy of structured priors in this section were also evaluated based on the proportion of the time they outperformed baseline priors in bias and variance. The calculated metric, share of simulations where the posterior median of structured priors outperformed the posterior median of baseline priors, are shown in figures referenced by the Improvement Proportion column of Appendix Table 1. A similar metric, the share of simulations where posterior variance of structured priors outperformed posterior variance of baseline priors, is referenced by the Improvement Proportion column of Appendix Table 2. In almost all cases, there was a strong improvement over the baseline priors more than half the time. This improvement was uniform across simulation scenarios, except for some simulations with sample size $n = 100$ and severe undersampling of the highest age groups.

When there is complex and structured variation of the true preference across the age categories, it is important for the model to not surpress this variation. Standard MRP estimates assume exchangeability between age categories and therefore shrink

cell estimates towards a global regression model in the absence of data. Structured priors for the age category covariate, while not perfectly recovering the true preference probabilities in the presence of limited data, are better able to capture much more of the variation in preference across age categories. This results in a better quantification of the variation across small areas.

When the number of categories for the structured covariates of interest is sufficiently large, they start to show beneficial effects on posterior MRP estimates. To quantify "sufficiently large" is problem-dependent as every structured prior will be different depending on the covariates of the data set. Furthermore, there are multiple structured priors one can choose from for a covariate. This is something we will not address here. We previously ran the same set of experiments in this results section for 3 and 6 age categories and did not observe a significant difference in posterior estimates for all three prior specifications. 12 age categories and more for our simulation studies are when the beneficial effects of structured priors become obvious.

4.2 Undirected Structured Priors Example: Spatial MRP

For the second simulation example, we will apply MRP for spatial data. More specifically, in the case of binomial regression to get subpopulation-level estimates for all 52 Public Use Microdata Areas (PUMA) in the state of Massachusetts. We simulate a nonrepresentative survey by purposefully oversampling certain spatial areas.

In this section, we will fit the models using the `R-INLA` package (Rue et al., 2009, 2017; Seppä et al., 2019), a fast approximate Bayesian inference package catered towards spatial modelling.

We will use penalized complexity (PC) prior specifications (Simpson et al., 2017) for hyperpriors in the spatial MRP model. The response variable for an observation is count data t_i and the total number of possible occurrences is T_i. Let n be the number of observations in the binomial response data set. Let $i = 1, \ldots, n$ be the index of counts. We define the following model below:

$$\begin{aligned}
t_i &\sim \text{Binomial}(\mu_i, T_i), \\
\mu_i &= \text{logit}^{-1}\left(\beta^0 + \alpha^{\text{PUMA}}_{j[i]} + \alpha^{\text{Education}}_{j[i]} + \alpha^{\text{Ethnicity}}_{j[i]}\right), \\
\alpha^{\text{Education}}_j | \sigma^{\text{Education}} &\sim \text{N}(0, (\sigma^{\text{Education}})^2), \text{ for } j = 1, \ldots, 6, \\
\alpha^{\text{Ethnicity}}_j | \sigma^{\text{Ethnicity}} &\sim \text{N}(0, (\sigma^{\text{Ethnicity}})^2), \text{ for } j = 1, \ldots, 6, \\
\sigma^{\text{Education}}, \sigma^{\text{Ethnicity}} &\sim \text{PC-Prior}(1, 0.1), \\
\beta^0 &\sim \text{N}(0, 1).
\end{aligned} \tag{4.6}$$

The poststratified estimand based on (4.6) for a subpopulation group $S \subseteq \{1, \ldots, J\}$ is then $\sum_{j \in S} \frac{N_j}{N} \mu_j$.

The *Besag-York-Mollié (BYM2) specification* models the prior spatial structure of PUMA effect, α^{PUMA}_j, as a BYM2 model (Morris et al., 2019; Riebler et al., 2016)

specified in (4.7). Let A_j be the set of first-degree neighbours for PUMA j. Let d_j be the cardinality of A_j. The BYM2 spatial prior for PUMA is defined as:

$$\begin{aligned} \alpha_j^{\text{PUMA}} &= \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\rho}\theta_j^* + \sqrt{\frac{\rho}{s}}\phi_j^*\right), \\ \phi_j^* &\sim \text{N}\left(\frac{\sum_{k\sim A_j}\phi_k^*}{d_j}, \frac{1}{d_j}\right), \\ \theta^* &\sim \text{N}(0, I_{52}), \end{aligned} \tag{4.7}$$

where we account for the multiple connected components using the method described in Freni-Sterrantino et al. (2018).

From (4.7), we see that ϕ^* is an ICAR prior and θ^* is an isotropic multivariate normal prior. The scaling factor s is computed so that the variance of $\sqrt{\frac{\rho}{\tau s}}\phi_j^*$ is approximately equal to 1 for all PUMA j. Additionally, we have the constraint $\sum_{j=1}^{52}\phi_j^* = 0$ since the joint distribution of ϕ^* is non-identifiable. We will specify $\tau \sim$ PC-Prior$(1, 0.1)$ (Simpson et al., 2017), which results in the precision τ to have density function $\frac{-\log(0.1)}{2}\tau^{-3/2}e^{\frac{\log(0.1)}{\tau^{1/2}}}$ and we'll specify the mixing parameter ρ being the default BYM2 hyperprior specification in INLA.

The *Independent and Identically Distributed (IID) specification* models the prior structure of PUMA effect, α^{PUMA}, as an isotropic multivariate normal distribution with precision $\tau \sim$ PC-Prior$(1, 0.1)$.

The IID specification does not allow for any borrowing of information. However, the BYM2 prior specification allows for posterior estimates of true preference for every PUMA to borrow information with its first-degree neighbours, where the amount of information borrowed is controlled by the posterior mixing hyperparameter ρ. This borrowing of information with first-degree neighbours aims to capture the smooth spatial structure of true preference across the 52 PUMA in Massachusetts, if there indeed is one.

Simulated Data

The sample We investigate the effects of over/undersampling 17 neighbouring PUMA areas near Boston. The true preference across 52 PUMA as well as the 17 cluster PUMA near Boston that are over/undersampled are visualized in Appendix Figure 31. The below simulations show that the spatial prior for PUMA effect, which is a BYM2 specification, outperforms the classical IID prior specification in MRP.

Assumed sample and population We will again assume that the population is sufficiently large so that sampling with replacement is equivalent to retrieving a random sample from the population.

Let $S \subseteq \{1, \ldots, J\}$ be the index set corresponding to the group of 17 PUMA near Boston. We perturb the probability of sampling an individual in S through 9 scenarios, ranging from 0.05 (under-representing individuals in S) to 0.82 (over-representing individuals in S).

True preferences for every poststratification cell $j \in \{1, \ldots, J\}$ in the population are generated with the following formula:

$$\mu_j = \text{logit}^{-1}\big(\beta^0 + X_{\text{PUMA}[j]} + X_{\text{Ethnicity}[j]} + X_{\text{Education}[j]}\big),$$

where $X_{\text{PUMA}[j]}$, $X_{\text{Ethnicity}[j]}$, $X_{\text{Education}[j]}$ corresponds to the PUMA, Race/Ethnicity and Education effect respectively for poststratification cell j. $X_{\text{PUMA}[j]}$, $X_{\text{Ethnicity}[j]}$, $X_{\text{Education}[j]}$ are defined in the appendix. M individuals with binary responses are then sampled from this poststratification matrix and then grouped based on poststratification cell to generate the binomial responses $(t_i, T_i)_{i=1}^n$. It's important to note that $X_{\text{PUMA}[j]}$ was generated through one random sample of a Gaussian Markov random field to ensure that the true preference μ_j has some degree of smoothness across all PUMA.

Undirected Structured Priors Results

Impact of prior choice on bias of posterior preferences We will again evaluate the impact of prior specification through perturbations of data regime for individuals near Boston. We can see from Figure 4 that the absolute bias for almost all 52 PUMA are significantly reduced when using a BYM2 specification instead of the IID specification, across all data regimes. The attenuation in absolute bias is apparent when the number of individuals sampled are 500 or 1000. This reduction in absolute bias is most apparent when individuals near Boston are over-represented (that is when the probability of response for S is close to 0.82). In this case, the BYM2 spatial prior outperforms the IID spatial prior by over 10 percent in absolute bias reduction. A spatial heatmap visualization of Figure 4 can be seen in Appendix Figure 32 and Appendix Figure 33. Based on these heatmaps, the PUMA that have the most absolute bias reduction when using the BYM2 spatial prior and when S is over-represented are the regions near the south-eastern coast of Massachusetts.

The share of simulations when the structured BYM2 spatial priors outperformed the IID priors for PUMA are shown in Appendix Figure 37. Differences in absolute bias from posterior medians and differences in posterior width were used as measures of comparison. The improvement was uniform over the IID priors, except for when there was severe oversampling of areas near Boston.

The average bias for every poststratification cell μ_j can be seen in Appendix Figure 34. At this granularity, the BYM2 spatial prior still outperforms the IID prior for PUMA for absolute bias reduction, regardless of sample size and what the probability of sampling S is.

Based off these spatial MRP simulations, we had also observed that the posterior width of every poststratification cell remains nearly the same when switching from the IID prior for PUMA to the BYM2 spatial prior for PUMA, as seen in Appendix Figure 35. This posterior width of every PUMA also remains nearly the same when switching from IID prior to BYM2 spatial prior, as seen in Appendix Figure 36.

In summary, structured priors for covariates with an undirected spatial structure are shown to improve MRP estimates through reduction of absolute bias. In this particular

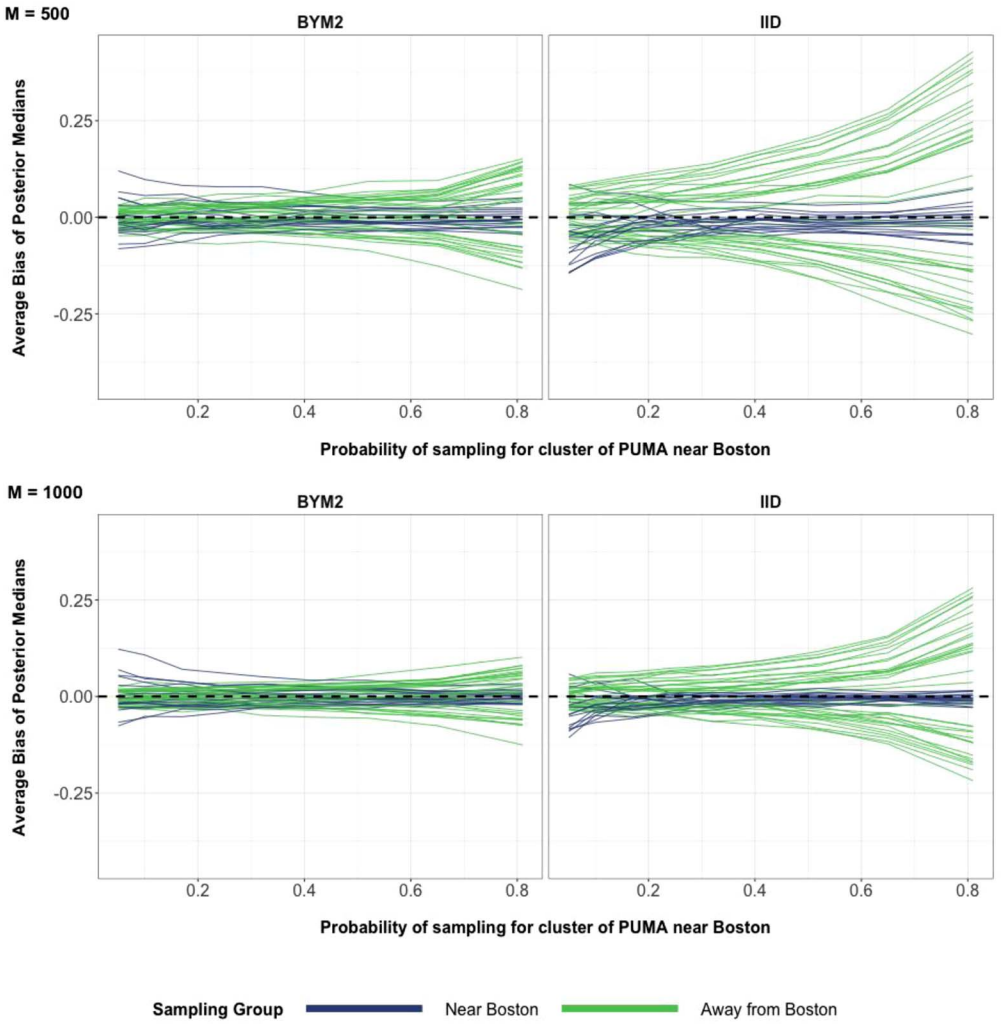

Figure 4: The average bias values coming from 200 simulations of poststratified estimates for the 52 PUMA areas. The possible values of average bias are in the interval $(-1, 1)$. M is the number of binary responses in every simulated data set. The top row corresponds to 500 binary responses used to define binomial responses for every simulation iteration. The bottom row corresponds to 1000 binary responses used to define binomial responses for every simulation iteration. The horizontal dashed line at y = 0 represents zero bias.

spatial MRP with count data simulation, we saw noticeable improvements with lower absolute bias for every PUMA when using the BYM2 spatial prior.

We had also run the same spatial MRP simulations but instead of having X_{PUMA} coming from a Gaussian Markov Random Field, we generated X_{PUMA} as a multivariate

independent normal. The BYM2 spatial prior on PUMA produced nearly the same posterior estimates as the IID spatial prior on PUMA, indicating that the BYM2 spatial prior does not force spatial structure when it's not present.

5 Analysis on U.S. Survey Data

Along with simulation studies that validate the benefit of structured priors, we further apply our approach to the National Annenberg Election Survey 2008 Phone Edition (NAES08-Phone) (Annenberg Center, 2008). NAES08-Phone was a phone survey conducted over the course of the 2008 US Presidential Election and the sampling methodology was based on random telephone number generation. NAES08-Phone observed a response rate of 23 percent. The population comes from the 2006–2010 5-year American Community Survey (ACS, United States Census Bureau / American FactFinder (2010)). The response variable of interest is whether an individual favors gay marriage or not. In 2008, this question was discussed heavily in the political landscape, as some states had not legalized same-sex marriage yet. The covariates used in the Annenberg survey sample are sex, race/ethnicity, household income, state of residence, age, education. The same covariates in the 5-year ACS are used so that poststratification and more specifically MRP can be performed.

5.1 National Annenberg Election Survey 2008

Table 1 contains the percentages of each factor for four of the covariates in the 2008 Annenberg phone survey and the 5-year ACS (excluding age and state of residence). A histogram summarizing the age covariate in the Annenberg phone survey is shown in the bottom plot of Figure 5. Individuals with ages 88 or above in the Annenberg phone survey had their ages set to 88. The reason for doing this was to maintain the non-identifiability of such individuals since they were the least frequently observed individuals in the Annenberg phone survey. The size of the Annenberg phone survey is 24,387 respondents.

Sex	Education	Race/Eth.	Income
Male: 43.3 (48.1)	Bachelor: 19.6 (16.9)	White: 80.2 (68.1)	Less than 10K: 4.8 (5.1)
Female: 56.7 (51.9)	Bachelor+: 19.0 (9.3)	Black: 8.6 (11.2)	[10K, 15K): 5.1 (4.1)
	Highschool: 28.5 (29.2)	Hispanic: 1.6 (5.2)	(15K, 25K]: 8.8 (9.1)
	No highschool: 7.1 (14.9)	Asian: 1.5 (4.8)	(25K, 35K]: 10.4 (9.7)
	Some college: 16.7 (22.5)	American Indian: 1.2 (1.0)	(35K, 50K]: 15.3 (14.0)
	2 year college: 9.0 (7.3)	Other: 7.0 (10.3)	(50K, 75K]: 19.1 (20.0)
			(75K, 100K]: 14.4 (13.9)
			(100K, 150K]: 12.3 (14.2)
			More than 150K: 9.9 (9.9)

Table 1: Percentage of each factor in the Annenberg phone survey and the 2006–2010 5-year American Community Survey (in parentheses) for sex, education, race/ethnicity, household income. Percentages are rounded to one decimal.

Appendix Figure 1 shows proportions of every state in the Annenberg survey and the differences in proportions between Annenberg survey and ACS. Ideally, there shouldn't be differences in the proportions for both surveys if both ACS and Annenberg survey are equally representative of the US population. Based on the bottom heatmap in Appendix Figure 1, California and Texas are shown to be the most under-represented by the Annenberg survey whereas New Hampshire and Missouri are the most over-represented by the Annenberg survey.

5.2 Poststratifying to the US Population

A smoothed density summarizing the age covariate in the ACS is shown in the bottom plot of Figure 5. The continuous age covariate in both the 5-year ACS and the Annenberg survey is discretized into either 12, 48, or 72 age categories in our analysis. In theory, the number of poststratification cells for Table 2 is $2 \times 6 \times 6 \times 9 \times 51 \times 78 = 2,577,744$. The cells left out by the expanded version of Table 2 are assumed have a population size of 0.

The 2006–2010 5-year ACS is a weighted probability survey, with a weight assigned to every individual in the sample. Based on the weights of individuals in the 5-year ACS, we form a 929,082-row poststratification matrix as seen in Table 2, which we will assume to be representative of the overall population for the 2008 Annenberg phone survey. We will use Table 2 to poststratify the 2008 Annenberg survey estimates to the US population. The ACS aggregates monthly probabilistic samples to form 1, 3, and 5-year ACS data sets. It aims to capture the most current demographic information annually, and answering the survey is mandatory. For these reasons, we believe that it's the most accurate representation of the US population every year.

Age	Sex	Race/Eth.	Education	State	Income	N
18	Male	American Indian	Highschool	Alabama	(15000, 25000]	35
18	Female	American Indian	Highschool	Arizona	Less than 10000	124
⋮	⋮	⋮	⋮	⋮	⋮	⋮
95	Female	White	2 year college	Rhode Island	(100000, 150000]	15

Table 2: Full poststratification matrix for the 5-year American Community Survey.

5.3 The Models for the 2008 Annenberg Phone Survey

For ages 18–40, the smooth ACS density in Figure 5 is higher than the Annenberg histogram, implying that the Annenberg survey underrepresents younger individuals. For the other demographic traits, relative to the 5-year ACS, Table 1 show that the Annenberg survey overrepresents white individuals and women.

Let $y_i = 1$ if respondent i favors same-sex marriage. Then we model,

$$\begin{aligned}\Pr(y_i = 1) = \text{logit}^{-1}\Big(&\beta^0 + \alpha_{j[i]}^{\text{Age Cat.}} + \beta^{\text{Sex}} X_{\text{Sex},i} + \alpha_{j[i]}^{\text{Ethnicity}} \\ &+ \alpha_{j[i]}^{\text{Education}} + \alpha_{j[i]}^{\text{State}} + \alpha_{j[i]}^{\text{Income}}\Big).\end{aligned} \tag{5.1}$$

We define the *baseline, autoregressive and random walk specifications* to have in common the priors distributions defined in Appendix Equation (0.1).

Let $J^{\text{Age Cat.}}$ be the number of categories for the continuous covariate age. The *baseline specification* has the prior distributions for age category:

$$\begin{aligned} \alpha_j^{\text{Age Cat.}} \mid \sigma^{\text{Age Cat.}} &\overset{\text{ind.}}{\sim} \text{N}(0, (\sigma^{\text{Age Cat.}})^2), \text{ for } j = 1, \ldots, J_{\text{Age Cat.}}, \\ \sigma^{\text{Age Cat.}} &\sim \text{N}_+(0, 1). \end{aligned} \tag{5.2}$$

The *autoregressive specification* has the prior distributions:

$$\begin{aligned} \alpha_1^{\text{Age Cat.}} \mid \rho, \sigma^{\text{Age Cat.}} &\sim \text{N}(0, \frac{1}{1-\rho^2}(\sigma^{\text{Age Cat.}})^2), \\ \alpha_j^{\text{Age Cat.}} \mid \alpha_{j-1}^{\text{Age Cat.}}, \ldots, \alpha_1^{\text{Age Cat.}}, \rho, \sigma^{\text{Age Cat.}} &\sim \text{N}(\rho\alpha_{j-1}^{\text{Age Cat.}}, (\sigma^{\text{Age Cat.}})^2), \\ &\text{for } j = 2, \ldots, J^{\text{Age Cat.}}, \\ \sigma^{\text{Age Cat.}} &\sim \text{N}_+(0, 1), \\ (\rho + 1)/2 &\sim \text{Beta}(0.5, 0.5). \end{aligned} \tag{5.3}$$

The *random walk specification* has the prior distributions:

$$\begin{aligned} \alpha_j^{\text{Age Cat.}} \mid \alpha_{j-1}^{\text{Age Cat.}}, \ldots, \alpha_1^{\text{Age Cat.}}, \sigma^{\text{Age Cat.}} &\sim \text{N}(\alpha_{j-1}^{\text{Age Cat.}}, (\sigma^{\text{Age Cat.}})^2), \\ &\text{for } j = 2, \ldots, J_{\text{Age Cat.}}, \\ \sum_{j=1}^{J_{\text{Age Cat.}}} \alpha_j^{\text{Age Cat.}} &= 0. \end{aligned} \tag{5.4}$$

We treat age as an ordered categorical predictor. It is reasonable to believe that people of similar ages will have similar attitudes on same-sex marriage. Hence we propose autoregressive and random walk structures as the prior distributions for age category.

5.4 Performing MRP with Structured Priors for the 2008 Annenberg Phone Survey

Hierarchical logistic regression with the two structured prior specifications and the baseline specification described previously are fit to the 2008 Annenberg phone survey. The poststratification matrix formed by the 5-year ACS is then used to poststratify posterior estimates for every age category. This is shown in Figure 5.

When age is discretized into 12 categories, there are no noticeable differences among the three prior specifications for age categories 1–11. Only at age category 12 do we start seeing a difference between the baseline specification and the two structured prior specifications. As expected, this difference in posteriors is observed when the underlying age category is a sparse cell for the survey data set. When age is discretized into 48 and 72 categories, one starts to see differences between the structured prior specifications and the baseline specification in terms of posterior variance for every age category.

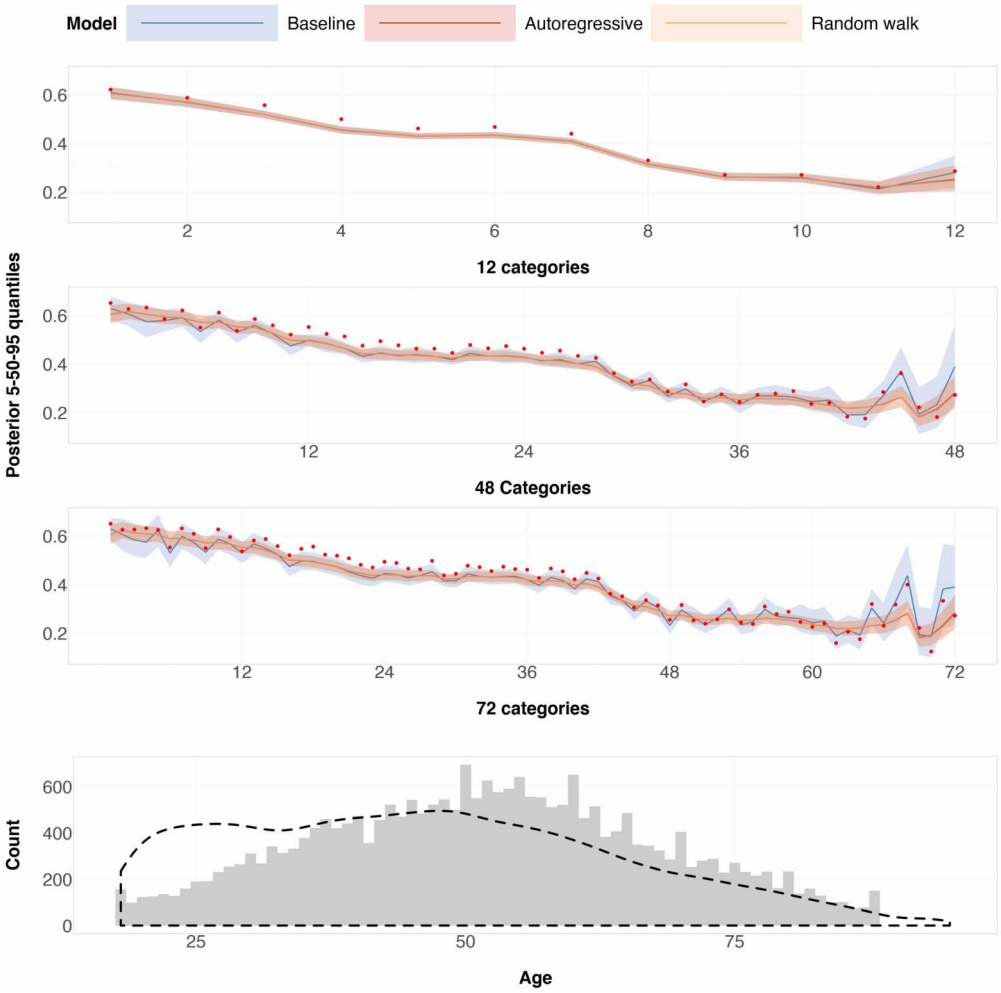

Figure 5: 12 (top), 48 (middle) and 72 (bottom) age categories. Red points in the top three plots are the empirical mean. The upper and lower bands in the top three plots correspond to the 95-percent and 5-percent posterior quantiles for every age category, and the middle solid line contains the posterior median for every age category. The density plot of ages in the ACS are coming from a random sample based off the 5-year ACS, where sampling is conducted with replacement using person weights given by the ACS. This random sample size is the same size as the 2008 Annenberg phone survey, and is assumed to be representative of the overall population defined by the 5-year ACS. 2000 iterations for 4 chains were run, for each prior specification and for age discretized into 12, 48 and 72 categories. The burn-in was set to 50 percent.

Posterior variances for the baseline specification are wider based on the 5-95 percent quantiles, and they expand a significant amount for the oldest age categories.

The baseline prior specification's posteriors become contracted towards their respective empirical means, which is not ideal since the empirical means swing more wildly for the older age categories. On the other hand, the autoregressive and random walk specifications are more smooth due to their property of having neighboring posterior random effects for age categories sharing information, and this is most noticeable when the number of age categories is 72. This smoothing effect is desirable for ordinal data as one may be interested in capturing a long-term trend when age increases.

What's also worth noting is that the baseline specification drastically changes the posterior variances when the number of categories for age changes from 12 to 48 to 72. Structured priors provide some stability in posterior variances despite how the input survey data is preprocessed through discretization of continuous variables.

The posterior population preferences for all three prior specifications remain nearly identical across the three age categories. This remains consistent with population preference results based on the simulation studies.

Based on the simulation studies, we had shown that structured priors reduce absolute bias and posterior variances of structured covariates. In our application of structured priors to the non-representative 2008 Annenberg phone survey, we see that structured priors reduce posterior variance on the structured covariate age as well.

6 Conclusion

We proposed using priors that exploit underlying structure in the covariates of multilevel regression and poststratification. Defined as structured prior distributions, they aim to introduce more intelligent shrinkage of posterior estimates.

We found through simulation studies that structured priors, when compared to independent random effects, reduce posterior MRP bias regardless of nonresponse pattern if there is an underlying pattern. A secondary benefit of structured priors when compared to independent random effects is that they reduce posterior variances for MRP estimates at the subpopulation levels corresponding to structured covariates of interest. We found that structured priors weather even extreme nonresponse patterns when compared to traditional random effects used in MRP. This is as expected since structured priors enable intelligent information-borrowing and shrinkage in posterior MRP estimates. This allows for an improved quantification of variation across small areas. Our modeling strategy of using structured priors was also applied to the non-representative 2008 Annenberg phone survey. The structured priors we describe here have similar smoothing properties to nonparametric regression methods such as GP regression and kernel smoothing (Gelman et al., 2013; Rasmussen, 2003).

Our investigations of using MRP for the Annenberg survey used ACS data to its full capacity through the usage of a 5-year ACS that covered the year 2008. Using a 1-year or a 3-year ACS which would have resulted in rougher information about the

population. Indeed, the information used to build the poststratification can be a limiting factor for MRP. The accuracy of poststratification in MRP is dependent on whether the poststratification matrix used is a true representation of the target population or not.

Based on both simulation studies and analysis on the Annenberg survey, we saw that more age categories resulted in lower posterior variance and bias for age category estimates. This comes at the tradeoff of coarser information about N_j, the size of the poststratification cell l. Another limitation one may have is deciding covariates to impose structured priors on. This choice is dependent on the modeller's knowledge of the problem and the data used.

There is usually more than one set of structured priors to propose, and this model selection and comparison problem is not addressed in this paper. The method in the paper could also be extended to using structured priors on interaction terms (Ghitza and Gelman, 2013), which is another active area in MRP research. Furthermore, we do not analyze the scenario when a structured prior is used for a covariate with no apparent structure.

In this manuscript we demonstrate improvements to MRP estimates through the use of structured priors when justified to do so. We believe that this is a contribution to the wider field considering other forms of regularization with MRP, but rather than employing black-box methods, using structured priors exploits methodologist and survey administrator knowledge.

Supplementary Material

Supplementary material for "Improving multilevel regression and poststratification with structured priors" (DOI: 10.1214/20-BA1223SUPP; .pdf).

References

Annenberg Center (2008). "The Annenberg Public Policy Center's National Annenberg Election Survey 2008 Phone Edition (NAES08-Phone) *[Data file and code book]*." Available from https://www.annenbergpublicpolicycenter.org/tag/data-sets/. 736

Besag, J. (1975). "Statistical analysis of non-lattice data." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3): 179–195. 723

Bisbee, J. (2019). "BARP: Improving Mister P Using Bayesian Additive Regression Trees." *American Political Science Review*, 113(4): 1060–1065. 719, 723

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1). 727

Chen, T. and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794. 723

Deming, W. E. and Stephan, F. F. (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *The Annals of Mathematical Statistics*, 11(4): 427–444. MR0003527. doi: https://doi.org/10.1214/aoms/1177731829. 720

Downes, M., Gurrin, L. C., English, D. R., Pirkis, J., Currier, D., Spittal, M. J., and Carlin, J. B. (2018). "Multilevel Regression and Poststratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples." *American journal of epidemiology*, 187(8): 1780–1790. 720

Elliott, M. R., Valliant, R., et al. (2017). "Inference for nonprobability samples." *Statistical Science*, 32(2): 249–264. MR3648958. doi: https://doi.org/10.1214/16-STS598. 720

Freni-Sterrantino, A., Ventrucci, M., and Rue, H. (2018). "A note on intrinsic conditional autoregressive models for disconnected graphs." *Spatial and spatio-temporal epidemiology*, 26: 25–34. 733

Gao, Y., Kennedy, L., Simpson, D., and Gelman, A. (2020). "Supplementary material for "Improving multilevel regression and poststratification with structured priors"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1223SUPP. 727

Gelman, A. (2018). "Regularized prediction and poststratification: A generalization of Mister P." *Statistical Modeling, Causal Inference, and Social Science Blog*. URL https://statmodeling.stat.columbia.edu/2018/05/19/regularized-prediction-poststratification-generalization-mister-p/ 719, 723

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press. 722

Gelman, A., Lax, J., Phillips, J., Gabry, J., and Trangucci, R. (2016). "Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion." *Unpublished manuscript, Columbia University*. 720

Gelman, A. and Little, T. C. (1997). "Poststratification into many categories using hierarchical logistic regression." 720, 721

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC. MR3235677. 740

Ghitza, Y. and Gelman, A. (2013). "Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups." *American Journal of Political Science*, 57(3): 762–776. 720, 741

Goplerud, M., Kuriwaki, S., Ratkovic, M., and Tingley, D. (2018). "Sparse Multilevel Regression (and Poststratification [sMRP])." *Unpublished manuscript, Harvard University*. 723

Izrael, D., Battaglia, M. P., and Frankel, M. R. (2009). "Extreme survey weight adjustment as a component of sample balancing (aka raking)." In *Proceedings from the Thirty-Fourth Annual SAS Users Group International Conference*. 720

Lax, J. R. and Phillips, J. H. (2009). "Gay rights in the states: Public opinion and policy responsiveness." *American Political Science Review*, 103(3): 367–386. 720, 724

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423–498. MR2853727. doi: https://doi.org/10.1111/j.1467-9868.2011.00777.x. 723

Little, R. J. (1993). "Post-stratification: a modeler's perspective." *Journal of the American Statistical Association*, 88(423): 1001–1012. 720

Lohr, S. L. (2009). *Sampling: design and analysis*. Nelson Education. MR3057878. 720

Morris, G. E. (2019). "If everyone had voted, Hillary Clinton would probably be president." *Economist*. URL https://www.economist.com/graphic-detail/2019/07/06/if-everyone-had-voted-hillary-clinton-would-probably-be-president 720

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., and DiMaggio, C. (2019). "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan." *Spatial and spatio-temporal epidemiology*, 31: 100301. 732

Park, D. K., Gelman, A., and Bafumi, J. (2004). "Bayesian multilevel estimation with poststratification: state-level estimates from national polls." *Political Analysis*, 12(4): 375–385. 720

Pfeffermann, D. et al. (2013). "New important developments in small area estimation." *Statistical Science*, 28(1): 40–68. MR3075338. doi: https://doi.org/10.1214/12-STS395. 719

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ 721

Rao, J. N. (2014). "Small-Area Estimation." *Wiley StatsRef: Statistics Reference Online*, 1–8. MR1953089. doi: https://doi.org/10.1002/0471722189. 719

Rasmussen, C. E. (2003). "Gaussian processes in machine learning." In *Summer School on Machine Learning*, 63–71. Springer. 740

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). "An intuitive Bayesian spatial model for disease mapping that accounts for scaling." *Statistical methods in medical research*, 25(4): 1145–1165. MR3541089. doi: https://doi.org/10.1177/0962280216660421. 732

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC. MR2130347. doi: https://doi.org/10.1201/9780203492024. 721, 723, 725

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2): 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 732

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). "Bayesian computing with INLA: a review." *Annual Review of Statistics and Its Application*, 4: 395–421. MR3634300. doi: https://doi.org/10.1214/16-STS576. 732

Seppä, K., Rue, H., Hakulinen, T., Läärä, E., Sillanpää, M. J., and Pitkäniemi, J. (2019). "Estimating multilevel regional variation in excess mortality of cancer patients using integrated nested Laplace approximation." *Statistics in medicine*, 38(5): 778–791. MR3916688. doi: https://doi.org/10.1002/sim.8010. 732

Si, Y., Trangucci, R., Gabry, J. S., and Gelman, A. (2017). "Bayesian hierarchical weighting adjustment and survey inference." *arXiv preprint* arXiv:1707.08220. 721

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 723, 732, 733

Skinner, C., Wakefield, J., et al. (2017). "Introduction to the design and analysis of complex survey data." *Statistical Science*, 32(2): 165–175. MR3648953. doi: https://doi.org/10.1214/17-STS614. 720

Stan Development Team (2016). "RStan: the R interface to Stan." *R package version*, 2(1). 727

Trangucci, R., Ali, I., Gelman, A., and Rivers, D. (2018). "Voting patterns in 2016: Exploration using multilevel regression and poststratification (MRP) on pre-election polls." *arXiv preprint* arXiv:1802.00842. 720, 724

United States Census Bureau / American FactFinder (2010). "2006–2010 ACS 5-year Public Use Microdata Samples (PUMS) – CSV format *[Data file and code book]*." Available from https://factfinder.census.gov/. 736

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). "Forecasting elections with non-representative polls." *International Journal of Forecasting*, 31(3): 980–991. 719, 720

Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J., and Croft, J. B. (2014). "Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system." *American journal of epidemiology*, 179(8): 1025–1033. 719, 720

Acknowledgments

Daniel Simpson and Yuxiang Gao were funded by the Canadian Natural Sciences and Engineering Research Council and the Canadian Research Chairs program. Andrew Gelman was supported by the U.S. Office of Naval Research, the Institute for Education Sciences and the National Science Foundation. We thank Shira Mitchell for helpful feedback. We also thank three reviewers and the editor for insightful feedback.