COMPARISON OF UNIT-LEVEL SMALL AREA ESTIMATION MODELING APPROACHES FOR SURVEY DATA UNDER INFORMATIVE SAMPLING

PAUL A. PARKER*
RYAN JANICKI
SCOTT H. HOLAN

Unit-level modeling strategies offer many advantages relative to the area-level models that are most often used in the context of small area estimation. For example, unit-level models aggregate naturally, allowing for estimates at any desired resolution, and also offer greater precision in many cases. We compare a variety of the methods available in the literature related to unit-level modeling for small area estimation. Specifically, to provide insight into the differences between methods, we conduct a simulation study that compares several of the general approaches. In addition, the methods used for simulation are further illustrated through an application to the American Community Survey.

KEYWORDS: Bayesian analysis; Informative sampling; Pseudo-likelihood; Small area estimation; Survey sampling.

PAUL A. PARKER is an Assistant Professor in the Department of Statistics, University of California Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA. RYAN JANICKI is a Principal Researcher with the U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA. SCOTT H. HOLAN is Professorin the Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100, USA and U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA.

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF Grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program, and NSF grants SES-1853096, NCSE-2215168, and NCSE-2215169. Support for this research at the Missouri Research Data Center (MURDC), through the University of Missouri Population, Education and Health Center Doctoral Fellowship, as well as through the U.S. Census Bureau Dissertation Fellowship Program is also gratefully acknowledged. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the NSF or U.S. Census Bureau. The DRB approval number for this paper is CBDRB-FY21-331.

*Address correspondence to Paul A. Parker, Department of Statistics, University of California Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA; E-mail: paulparker@ucsc.edu.

https://doi.org/10.1093/jssam/smad022
Published by Oxford University Press on behalf of the American Association for Public Opinion Research 2023.

Statement of Significance

This manuscript provides an empirical comparison of common unit-level modeling approaches for small area estimation in the presence of informative sampling. Specifically, through the use of American Community Survey data, we conduct an empirical simulation study that allows for comparison between methods. We also apply a variety of methodologies to the important problem of poverty estimation using the American Community Survey.

1. INTRODUCTION

There is a growing demand for population estimates at a granular level based on survey data. In many cases, sample sizes for these granular estimates are too small to yield reliable results when considering traditional design-based survey estimators. Small area estimation models are critical in this landscape, in order to borrow strength from neighboring areas/domains, resulting in more precise estimates.

Models for small area estimation (SAE) can generally be thought of as falling into one of two categories: area level or unit level. Area-level models are fit using the design-based direct estimates, whereas unit-level models are fit using the individual survey responses. Our focus herein is on unit-level modeling strategies. Modeling individual survey responses comes with a number of challenges. For example, data sizes are typically much larger at the unit level. In addition, survey designs may often result is dependence between the unit probabilities of selection and the response of interest, even after conditioning on available auxiliary and design variables (Eideh and Nathan 2009). This scenario is termed informative sampling, and can lead to biased estimates when left unaccounted for.

Unit-level models also offer many potential advantages over area-level models. For example, they may yield more precise estimates (Hidiroglou and You 2016). They also allow for use of unit-level covariates such as demographic variables. Another advantage is that they can offer internal consistency when constructing estimates at different resolutions.

Parker et al. (2023) provide a recent review of the various unit-level modeling approaches within SAE that account for informative sampling designs. The methods discussed in this review are general, and not specific to certain data types. However, many important survey variables at the unit level tend to be binary or categorical in nature. For example, Parker et al. (2022) develop a unit-level model for categorical data that is applied to the problem of health insurance estimation similar to the estimates produced by the Small Area Health Insurance Estimates (SAHIE) Program (Luery 2011; Bauder et al. 2018). Poverty is another important binary variable that is relevant to the Small Area Income and Poverty Estimates (SAIPE) program (Bell et al. 2016). Even more recently, Sun et al. (2022) use unit-level models for binary data to produce estimates of expected job loss by state during the COVID-19 pandemic through the use of data from the Household Pulse Survey.

We note that this paper builds on the overview given by Parker et al. (2023). Specifically, the aim of this paper is to evaluate a selection of the strategies reviewed by Parker et al. (2023). Motivated by the American Community Survey (ACS) and other complex surveys, this is done by fitting different unit-level models on both simulated data, and on real ACS confidential micro-data, thereby comparing model-based predictions and uncertainty estimates. In addition to this, we provide some additional review of unit-level models that are specific to the case of binary survey data for completeness. We mainly use Bayesian methods for inference, but note that many model-based methods are general enough to be implemented in either setting. In the simulation studies and data examples given in sections 3 and 4, we fit three unit-level small area Bayesian models, with vague, proper priors on all unknown model parameters. Inference on the finite population parameters of interest is done using the posterior mean as a point estimate, and the posterior variance and credible intervals as measures of uncertainty. Importantly, in this paper we are not attempting to unify frequentist and Bayesain approaches in the context of finite population inference, though this is an interesting area of research. Related to this topic of research is the work of Little (2012) on *Calibrated Bayes* (cf. Parker et al. 2023).

The remainder of this paper is organized as follows. In section 2 we briefly review some of the methods available to account for informative sampling, noting that an in-depth review of these methods (as well as others) is given by Parker et al. (2023). In section 3 we compare three selected general models to a direct estimator under a simulation study designed around ACS data. Specifically, this simulation examines three Bayesian methods that span different general modeling approaches (pseudo-likelihood, nonparametric regression on the weights, and differing sample/population likelihoods) with the goal of examining the utility of each approach. For a thorough review of each of these approaches, see Parker et al. (2023). The Stan code used to fit these models is available at https://github.com/paparker/Unit_Level_Models. Similarly, section 4 uses the same models for a poverty estimates application similar to the SAIPE program. Finally, we provide concluding remarks in section 5.

2. BACKGROUND

The pseudo-likelihood (PL) approach provides a flexible framework for handling complex survey data under informative sampling. To account for the informative sampling mechanism, the likelihood is exponentially weighted by the survey weight for each unit in the sample

$$\mathrm{PL}(\boldsymbol{y}|\theta) = \prod_{i \in \mathcal{S}} f(y_i|\theta)^{w_i}.$$

Multiplying the PL by a suitable prior density, $\pi(\boldsymbol{\theta})$, Savitsky and Toth (2016) show that this leads to a valid pseudo-posterior distribution in a Bayesian model

$$\widehat{\pi}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w}) \propto \left\{ \prod_{i \in \mathcal{S}} f(y_i|\theta)^{w_i} \right\} \pi(\boldsymbol{\theta}),$$

provided the survey weights, w_i, are scaled to sum to the sample size, n.

Another approach to account for informative sampling is to regress the response variable on a nonlinear function of the weights. Such an approach requires prediction of the survey weights for population units outside of the sample. In this light, Vandendijck et al. (2016) extend the work of Si et al. (2015) in the context of SAE. For the proposed method, the authors assume that the poststratification cells are designated by the unique weights and scaled to sum to the sample size within each area. Then, similar to Si et al. (2015), a multinomial model is used to conduct poststratification using the posterior distribution. Assuming a Bernoulli response, they use the data model

$$\begin{aligned} y_{ij}|\eta_{ij} &\sim \mathrm{Bernoulli}(\eta_{ij}) \\ \mathrm{logit}(\eta_{ij}) &= \beta_0 + \mu(\tilde{w}_{ij}) + u_i + v_i \end{aligned}$$

for unit j in small area i, with $\tilde{w}_{ij}$ denoting the area-specific scaled survey weights. Independent area level random effects are given by v_i, whereas u_i denotes spatially dependent area-level intrinsic conditional autoregressive (ICAR) random effects (Besag 1974). Lastly, the authors explore the use of a Gaussian process prior over the function $\mu(\cdot)$ and use a multinomial model for poststratification.

One further approach to unit-level modeling in the presence of informative sampling is to infer a model for unsampled units based on a specified population model as well as a model for the survey weights. For example, suppose the finite population values y_{ij} are independent realizations from a population with density $f_p(\cdot|\boldsymbol{x}_{ij}, \boldsymbol{\theta})$, conditional on a vector of covariates $\boldsymbol{x}_{ij}$, and model parameters $\boldsymbol{\theta}$. Define the sample density, f_s (Pfeffermann et al. 1998) as the density function of y_{ij}, given that y_{ij} has been sampled

$$f_s\left(y_{ij}|\boldsymbol{x}_{ij},\boldsymbol{\theta},\gamma\right) = f_p\left(y_{ij}|\boldsymbol{x}_{ij},\boldsymbol{\theta},I_{ij}=1\right) = \frac{P\left(I_{ij}=1|y_{ij},\boldsymbol{x}_{ij},\gamma\right)f_p\left(y_{ij}|\boldsymbol{x}_{ij},\boldsymbol{\theta}\right)}{P\left(I_{ij}=1|\boldsymbol{x}_{ij},\gamma\right)},$$

where I_{ij} is a binary variable indicating whether unit j in area i has been sampled and γ is a vector of sampling model related parameters.

Pfeffermann and Sverchkov (2007) adapted the sample distribution to multilevel models for SAE of finite population means when both small areas and units within small areas are sampled with unequal probabilities in a two-stage, informative survey design, and showed how small area means can be predicted using the observed unit level data under a multilevel, informative survey design. In addition, they showed that the model for the survey weights can be specified conditionally on the response variables to account for the informativeness of the survey design. Possible models for the sample weights considered in the literature include linear and exponential models. Pfeffermann and Sverchkov (2007) considered the case of continuous response variables, y_{ij}, and modeled the sampled response data using the nested error regression model.

3. SIMULATION STUDY

Unit-level models offer several potential benefits (e.g., no need for benchmarking (Battese et al. 1988) and increased precision (Hidiroglou and You 2016)), however, accounting for the informative design is critical at the unit level. There are a variety of ways to approach this; however, the utility of each approach is not apparent. We choose three methods that span different general modeling approaches (pseudo-likelihood, nonparametric regression on the weights, and differing sample/population likelihoods), in order to address this question. We choose to sample a population based on existing survey data from a complicated design, and make estimates for poverty (similar to SAIPE).

To construct a simulation study, we require a population for which the response is known for every individual, in order to compare any estimates to the truth. It is also desirable to have an informative sample. We treat the 2014 ACS sample from Minnesota as our population (around 120,000 observations and 87 counties), and further sample 10,000 observations in order to generate our estimates from the selected models. Ideally, we would mimic the survey design used by ACS, however the design is highly complex which makes replication difficult. Instead, we subsample the ACS sample with probability proportional to the reported sampling weights, $w_{ij}^{(o)}$, using the Midzuno method (Midzuno 1951) from the `sampling` package in `R` (Tillé and Matei 2016).

This results in a new set of survey weights $w_{ij}^{(n)}$, which are inversely proportional to the original weights given in the ACS sample. Sampling in this manner results in a sample for which the selection probabilities are proportional to the original sampling weights. Sampling was done in this way to induce an informative subsample. By comparing weighted and unweighted direct estimates (not reported), we verify that sampling in this way yields an informative sample. Based on the models discussed in Parker et al. (2023), we fit three models to the newly sampled dataset, and create county level estimates of the proportion of the original ACS sample below the poverty level.

The first model we consider, which we call model 1, incorporates the survey design through a Bernoulli pseudo-likelihood with a logit link function.

3.1 Model 1

$$\begin{gathered} y_{ij}|\boldsymbol{\beta}, \boldsymbol{\mu} \propto \text{Bernoulli}(y_{ij}|p_{ij})^{\tilde{w}_{ij}} \\ \text{logit}(p_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i \\ u_i \overset{i.i.d.}{\sim} \text{N}(0, \sigma_u^2) \\ \boldsymbol{\beta} \sim \text{N}_p(\boldsymbol{0}_p, \boldsymbol{I}_{p\times p}\sigma_\beta^2), \quad \sigma_u \sim \text{Cauchy}^+(0, \kappa_u), \end{gathered} \tag{1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$ and the weights $\tilde{w}_{ij}$ are scaled to sum to the total sample size, so that $\sum_{i,j} \tilde{w}_{ij} = n$, as recommended by Savitsky and Toth (2016). Note that $\text{Bernoulli}(y_{ij}|p_{ij})^{\tilde{w}_{ij}}$ represents a Bernoulli probability mass function with probability p_{ij} evaluated at the data point y_{ij} and exponentiated by the power $\tilde{w}_{ij}$. $\text{Cauchy}^+(0, \kappa_u)$ denotes a half-Cauchy distribution, with scale parameter κ_u (Gelman 2006). We incorporate a vague prior distribution by setting $\sigma_\beta^2 = 10$ and $\kappa_u = 5$. This approach is based on the Bayesian pseudo-likelihood given in Savitsky and Toth (2016), where each Bernoulli likelihood contribution is exponentiated according to the scaled survey weight $\tilde{w}_{ij}$ in the first line of (1). The model structure is similar to that of Zhang et al. (2014), although we use the psuedo-likelihood in a Bayesian context rather than a frequentist one. Our design matrix $\boldsymbol{X}$ includes terms for age category, race category, and sex. We use poststratification, where the poststratification cells are defined by all cross classifications of age, race, and sex categories, by generating the nonsampled population at every iteration of our MCMC, which we use to produce our estimates. The poststratification cells consist of the unique combinations of county, age category, race category, and sex, for which the population sizes are known to us. See Parker et al. (2022) and the references therein for detailed discussion on poststratification.

The second model, labeled model 2 below, uses an unweighted Bernoulli likelihood with logit link function, but regresses on the survey weights to account for the informative survey design.

3.2 Model 2

$$
\begin{aligned}
y_{ij}|\beta_0, f(w_{ij}), \boldsymbol{u}, \boldsymbol{v} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit}(p_{ij}) &= \beta_0 + f(w_{ij}) + u_i + v_i \\
f(w_{ij})|\gamma, \rho &\sim \text{GP}(0, \text{Cov}(f(w_{ij}), f(w_{i'j'}))) \\
\text{Cov}(f(w_{ij}), f(w_{i'j'})) &= \gamma^2 \exp\left(-\frac{(w_{ij} - w_{i'j'})^2}{2\rho^2}\right) \\
\boldsymbol{u}|\tau, \alpha &\sim \text{N}(0, \tau \boldsymbol{D}(\boldsymbol{I} - \alpha \boldsymbol{W})^{-1}) \\
v_i|\sigma_v^2 &\sim \text{N}(0, \sigma_v^2), \quad i = 1, \ldots, m \\
\beta_0 \sim \text{N}(0, \sigma_\beta^2), \quad \gamma &\sim \text{Cauchy}^+(0, \kappa_\gamma), \quad \rho \sim \text{Cauchy}^+(0, \kappa_\rho) \\
\tau \sim \text{Cauchy}^+(0, \kappa_\tau), \quad \alpha &\sim \text{Unif}(-1, 1), \quad \sigma_v \sim \text{Cauchy}^+(0, \kappa_v),
\end{aligned} \tag{2}
$$

where $\boldsymbol{u} = (u_1, \ldots, u_D)'$ is a vector of spatially correlated random effects, $\boldsymbol{v} = (v_1, \ldots, v_D)'$ is a vector of independent random effects, $\boldsymbol{D}$ is a diagonal matrix containing the number of neighbors for each area $i = 1, \ldots, m$ and $\boldsymbol{W}$ is an area adjacency matrix. Again, we use a vague prior distribution by setting $\sigma_\beta^2 = 10$ and $\kappa_\gamma = \kappa_\rho = \kappa_\tau = \kappa_v = 5$. This is similar to the work of Vandendijck et al. (2016), but using the squared exponential covariance kernel as in Si et al. (2015), rather than a random walk prior on $f(\cdot)$. See also section 4 of Parker et al. (2023) which discusses inclusion of survey weights in a model. Additionally, we choose to use the conditional autoregressive (CAR) structure rather than ICAR structure on our random effects $\boldsymbol{u}$. The random effects $\boldsymbol{u}$ and $\boldsymbol{v}$ are included to allow for "borrowing strength" both globally and locally (Besag et al. 1991). Only the sum of the random effects and not their individual values is identifiable; however, the posterior will be proper so long as at least one of the prior distributions on the variance components is proper (Eberly and Carlin 2000). Note that although Vandendijck et al. (2016) use the weights scaled to sum to county sample sizes as inputs into the non-parametric function $f(\cdot)$, we attained better results by using the unscaled weights. We use the multinomial model

$$
(n_{1k}, \ldots, n_{L_k k}) \sim \text{Multinomial}\left(n_k; \frac{N_{1k}/w_{(1)k}}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}, \ldots, \frac{N_{L_k k}/w_{(L_k)k}}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}\right) \tag{3}
$$

to model the population weight values, in order to perform poststratification. In this model, n_{lk} represents the sample size in poststrata cell l in area k, while N_{lk} represents the population size in the same cell. Poststratification cells are determined by unique weight values within each county, denoted $w_{(l)k}$. Because all units in the same cell will share the same weight, by determining the population size of each cell, the weights are implicitly determined, and thus the population may be generated using the model specified in (2).

The final model we consider is labeled model 3 below and incorporates the effect of an informative design by regressing the log of the survey weights on the response variable.

3.3 Model 3

$$
\begin{aligned}
y_{ij}|p_{ij} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit}(p_{ij}) &= \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + u_i \\
\log(w_{ij})|y_{ij} &\sim \boldsymbol{x}_{ij}^T\boldsymbol{\alpha} + y_{ij} * a + \epsilon_{ij} \\
u_i &\overset{i.i.d.}{\sim} \text{N}(0, \sigma_u^2) \\
\epsilon_{ij} &\overset{i.i.d.}{\sim} \text{N}(0, \sigma_\epsilon^2),
\end{aligned}
$$

with vague $\text{N}(0, 10)$ priors on the regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and a, and vague $\text{Cauchy}^+(0, 5)$ priors on the variance components σ_u and σ_ϵ. This model acts as a Bayesian extension of Pfeffermann and Sverchkov (2007). Notably, in all three models, the prior specification is vague relative to the scale of the data and, therefore, imparts little impact on our analyses.

All three models were fit via HMC using Stan (Carpenter et al. 2017). We treat poverty status (0 for not in poverty or 1 for in poverty) as our response variable, and use age category, race category, and sex as covariates. These covariate values are known for the entire population, allowing for prediction of unsampled units. We ran each model using two chains, each of length 2,000, and discarding the first 1,000 iterations as burn-in, thus using a total of 2,000 MCMC samples. Convergence was assessed visually via traceplots of the sample chains, with no lack of convergence detected. We repeated the simulation 100 times, with a sample size of 10,000 each time. That is, we create 100 distinct subsamples from the ACS sample, and fit the three models to each subsample. Denoting $\widehat{Y}_{ir}$ as the estimate for county i in subsample r, and Y_i as the true value of the population parameter in county i, we compare the root mean squared error (RMSE)

$$
\sqrt{\frac{1}{m \times 100}\sum_{i=1}^{m}\sum_{r=1}^{100}(\widehat{Y}_{ir} - Y_i)^2}
$$

and absolute bias

$$\frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{100}\sum_{r=1}^{100}\widehat{Y}_{ir}-Y_i\right|.$$

Although we present bias, we mainly focus on prediction RMSE in our comparisons, as this metric simultaneously strikes a balance between bias and variance reduction. We also report the Kendall tau rank distance from the true county poverty rate rankings, and 95 percent credible interval coverage rate for county level estimates

$$\frac{1}{m\times 100}\sum_{i=1}^{m}\sum_{r=1}^{100}I(\widehat{L}_{ir}\leq Y_i\leq \widehat{U}_{ir}),$$

averaged over both counties and simulated datasets, as well as computation time in seconds for each model in table 1. Note that $\widehat{L}_{ir}$ and $\widehat{U}_{ir}$ represent the lower and upper 95 percent credible interval bounds respectively for county i and subsample r. These metrics are accompanied by their corresponding bootstrapped standard errors. We also compare to a Horvitz–Thompson (HT) direct estimator as well as to a survey weighted poststratification design-based estimator (WPS, see Lohr (2019), page 374 for details).

Each of the three model-based estimators provides a substantial reduction in RMSE compared to the direct estimator, with model 3 being the best in this regard. Additionally, Model 1 gives a low bias, quite comparable to the direct estimators. Note that the ratio of the squared bias to the MSE is roughly in the range of 2.5–6 percent for all procedures, except for model 3, which is approximately 55 percent. The coverage rates of the credible intervals are also reported

Table 1. Simulation Results: RMSE×10^{-2}, Bias×10^{-2}, Rank Distance×10^{3}, 95 Percent Credible Interval Coverage Rate, and Computation Time in Seconds were Averaged over 100 Simulations in order to Compare the Direct Estimator to Three Model Based Estimators and Unweighted Direct Estimate

Estimator	RMSE×10^{-2}	Bias×10^{-2}	Rank Dist.×10^{3}	CI Cov. Rate	Time (s)
HT	6.54 (1.2×10^{-3})	1.0 (*)	1.212 (1.0×10^{1})	NA	NA
WPS	6.55 (9×10^{-4})	1.7 (*)	1.194 (1.1×10^{1})	NA	NA
Model 1	4.17 (8×10^{-4})	1.0 (*)	1.183 (1.0×10^{1})	0.86	106
Model 2	4.10 (4×10^{-4})	1.0 (*)	1.120 (0.9×10^{1})	0.894	5,948
Model 3	2.96 (3×10^{-4})	2.2 (*)	1.031 (0.8×10^{1})	0.943	437

Note.—Standard errors are denoted in parentheses. Note that some standard errors were suppressed due to rounding requirements necessary for disclosure avoidance. These have been denoted with (*).

in table 1. Model 3 produces intervals that nearly exactly achieve the nominal 95 percent coverage rate, with models 1 and 2 falling below 90 percent. The reasons for the under coverage with models 1 and 2 are not entirely clear; it could be due to model misspecification or due to this particular survey design. Coverage rates for the design-based estimators were not reported as we do not have access to replicate weights or joint inclusion probabilities in this simulation study.

Model 1 requires substantially less computation time compared to the other model-based estimators, especially when comparing to model 2. This suggests that if one wanted to scale the model to include more data, such as estimates at a national level, model 1 may be easier to work with. Computation times will vary depending on the specific resources used, however the main focus here is the relative time between models. Additionally, this simulation illustrates that it is feasible to fit Bayesian unit-level models in practice under reasonable computation times.

In some cases, interest may not be in the specific county point estimates, but rather the relative ranking of these point estimates. Thus, it is desirable to select a model that has both lower RMSE and lower rank distance compared to the direct estimates. It is clear that each of the model based estimates is able to reduce this rank distance compared to the direct estimators, with model 3 performing exceptionally well in this case.

In figure 1, we show the average reduction in RMSE, for each county, that was attained by the three model based estimators when compared to the HT direct estimator, averaged over the 100 simulations. Counties that did not see a reduction are plotted in gray. Although model 2 had roughly 16 percent of counties that saw an increase in RMSE, this only occurred in counties that had lower RMSE for the direct estimate already, and the increases tended to be minimal. There are some important differences between the model results here.

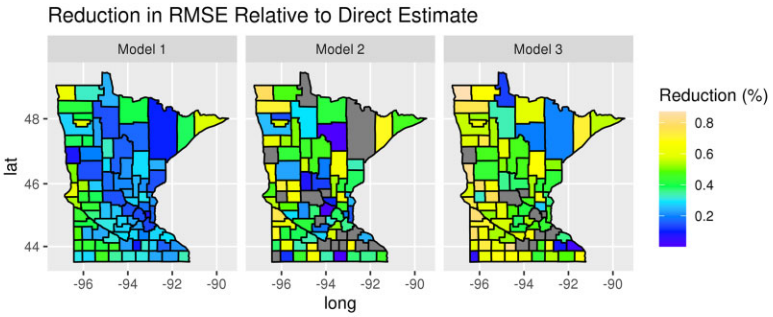

Figure 1. Model Reduction in RMSE Compared to the Direct Estimates, Averaged Over 100 Simulations. Counties that did not see a reduction are not plotted (shown in gray).

Specifically, model 1 achieves a reduction in nearly every county unlike the other two models, but model 3 tends to achieve a greater reduction in RMSE in general when compared to model 1.

4. POVERTY ESTIMATE DATA ANALYSIS

The SAIPE program is a U.S. Census Bureau program that produces estimates of median income and the number of people below the poverty threshold for states, counties, and school districts, as well as for various subgroups of the population. The SAIPE estimates are critical in order for the Department of Education to allocate Title I funds.

The current model used to generate SAIPE poverty estimates is an area-level Fay-Herriot model (Fay and Herriot 1979) on the log scale. The response variable is the log transformed HT direct estimates from the single year ACS of the number of individuals in poverty at the county level. The model includes a number of powerful county level covariates such as the number of claimed exemptions from federal tax return data, the number of people participating in the Supplemental Nutrition Assistance Program (SNAP), and the number of Supplemental Security Income (SSI) recipients. Luery (2011) provides a comprehensive overview of the SAIPE program, including the methodology used to produce various area-level estimates and the covariates used in the model.

We use a single year of ACS data (2014 again) from Minnesota to fit the three models described in section 3, using the same response and covariates. The model based estimators we present are not meant to replace the current SAIPE methodology, but rather to illustrate how unit-level models can be used in an informative sampling application such as this one. The model-based predictions of the proportion of people below the poverty threshold by county under each method are presented and compared with a direct estimator.

In figure 2, we show the estimate of the proportion of people below the poverty level by county for each of the model-based estimators as well as the HT direct estimator. Note that a small amount of noise has been added to the HT direct estimates as a disclosure avoidance practice. All of the estimates here seem to capture the same general spatial trend. The model based estimates resemble smoothed versions of the direct estimates, especially in the more rural areas of the state. Small sample sizes can lead to direct estimates with high variance, but the model based approaches can "share information" across areas, which leads to more precise estimates. We also compare the reduction in model based standard errors when compared to the HT direct estimate in figure 3. This illustrates the precision that is gained by using a model-based estimator rather than a direct estimator in an SAE setting. Model 3 in particular appears to have the lowest standard errors in more rural areas and model 1 seems to have lower standard errors in more populated areas. For this particular

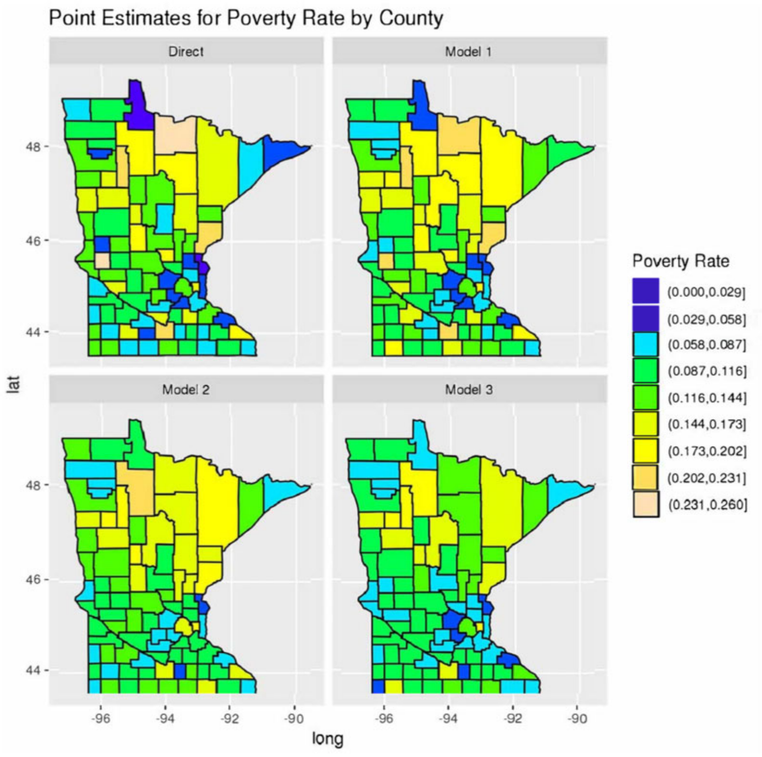

Figure 2. Noise Infused HT Direct and Model Based Point Estimates of Poverty Rate by County for Minnesota in 2014.

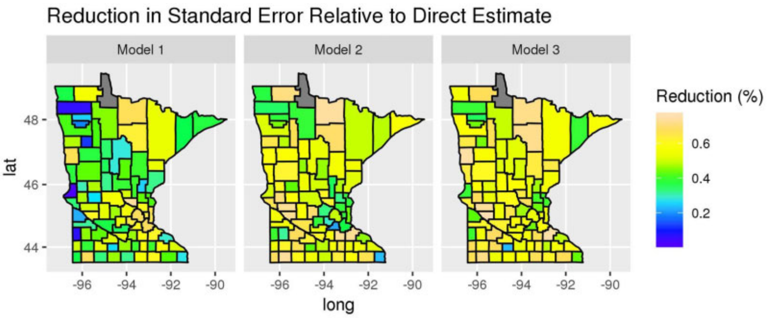

Figure 3. Model Based Reduction in Standard Errors for Poverty Rate by County. Counties that did not see a reduction are not plotted (shown in gray).

application, all three of the models we explored would be valid choices, with substantial reductions in RMSE as shown in section 3.

In this case, the population cell sizes were known, however in many applications they may not be, in which case model 2 would likely be the best option. In other cases where incorporating covariate information is desired, model 2 is not well equipped to make estimates, due to the multinomial model (3). This application was conducted for a single state, however if one wanted to scale the analysis, for example making estimates for every county in the United States, model 1 appears to be the most computationally efficient. An approach similar to model 2, albeit using a different nonlinear regression approach from the Gaussian Process regression considered here, may also be computationally efficient. Vandendijck et al. (2016) reported strong results using splines for this setup. Overall we found that each of these unit-level methods can offer precise area-level estimates, however, the properties of the particular dataset under consideration as well as the goals of the user should drive which model is selected.

Modeling poverty counts at the unit level has a number of benefits when compared to area-level models. Specifically, the current SAIPE model is on the log scale, and thus cannot naturally accommodate estimates for areas with a corresponding direct estimate of zero, whereas unit-level modeling need not be on the log scale, and thus does not suffer from this problem. Additionally, making predictions at multiple spatial resolutions is straightforward in the unit-level setting, as predictions can be generated for all units in the population and then aggregated as necessary, that is, the so-called bottom-up approach. Under a unit-level approach, one could generate poverty estimates at both a county level and school district level under the same model. Notably, spatial random effects in this setting could be placed at different levels of geography, though exact model specification is often problem specific (e.g., accounting for a nested or other spatial structure). In this case, this can be viewed as a multiscale model (Ferreira et al. 2011). In addition to these structural benefits, table 1 illustrates that unit-level models have the capacity to provide substantial reductions in MSE and variance when compared to direct estimators.

5. CONCLUSION

In the context of SAE, we have described several strategies for unit-level modeling under informative sampling designs and illustrated their effectiveness relative to design-based estimators (direct estimates). Specifically, motivated by the ACS and other complex surveys, our simulation study (section 3) illustrated three model-based estimators that exhibited superior performance relative to the direct estimator in terms of MSE, with model 3 performing best in this regard. Among the three models compared in this simulation, model 1 displayed the lowest computation time relative to the other model-based estimators and, therefore, may be advantageous in higher-dimensional settings. While this design-based simulation illustrated the superiority of model-based methods, it is important in future work to consider other simulation setups which include different degrees of informativeness as well as model and design-model (superpopulation) setups as a sensitivity check as well as to understand the reasons for the undercoverage of the interval estimates for some methods.

Although the focus here was not on optimizing computation, it is likely that tools such as INLA or variational approximations could further reduce the computing time necessary to fit these types of models. Along these same lines, spatial models such as model 2 may scale well considering a divide-and-conquer approach such as estimating counties within a state using only data from that state and its immediate spatial neighbors.

The models in section 3 (and section 4) constitute modest extensions to models currently in the literature. Specifically, model 2 provides an extension to Vandendijck et al. (2016), whereas model 3 can be seen as a Bayesian version of the model proposed by Pfeffermann and Sverchkov (2007).

REFERENCES

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), “An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data,” *Journal of the American Statistical Association*, 83(401), 28–36.

Bauder, M., Luery, D., and Szelepka, S. (2018), “Small Area Estimation of Health Insurance Coverage in 2010–2016.” Technical Report, Small Area Methods Branch, Social, Economic, and Housing Statistics Division, U.S. Census Bureau, Suitland, MD.

Bell, W. R., Basel, W. W., and Maples, J. J. (2016), “An Overview of the U. S. Census Bureau’s Small Area Income and Poverty Estimates Program,” in *Analysis of Poverty Data by Small Area Estimation*, ed. M. Pratesi, New York: Wiley, pp. 349–378.

Besag, J. (1974), “Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion),” *Journal of the Royal Statistical Society. Series B*, 36, 192–225.

Besag, J., York, J., and Mollié, A. (1991), “Bayesian Image Restoration, with Two Applications in Spatial Statistics,” *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

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, 1–32.

Eberly, L. E., and Carlin, B. P. (2000), “Identifiability and Convergence Issues for Markov Chain Monte Carlo Fitting of Spatial Models,” *Statistics in Medicine*, 19, 2279–2294.

Eideh, A., and Nathan, G. (2009), “Two-Stage Informative Cluster Sampling–Estimation and Prediction with Application for Small-Area Models,” *Journal of Statistical Planning and Inference*, 139, 3088–3101.

Fay, R. E., and Herriot, R. A. (1979), “Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data,” *Journal of the American Statistical Association*, 74(366a), 269–277.

Ferreira, M. A., Holan, S. H., and Bertolde, A. I. (2011), “Dynamic Multiscale Spatiotemporal Models for Gaussian Areal Data,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5), 663–688.

Gelman, A. (2006), “Prior Distributions for Variance Parameters in Hierarchical Models,” *Bayesian Analysis*, 1(3), 515–533.

Hidiroglou, M. A., and You, Y. (2016), "Comparison of Unit Level and Area Level Small Area Estimators," *Survey Methodology*, 42, 41–61.

Little, R. J. (2012), "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics," *Journal of Official Statistics*, 28(3), 309.

Lohr, S. L. (2019), *Sampling: Design and Analysis*, Boca Raton, FL: Chapman and Hall/CRC.

Luery, D. M. (2011), "Small Area Income and Poverty Estimates Program." in Proceedings of 27th SCORUS Conference, pp. 93–107, Jurmala, Latvia.

Midzuno, H. (1951), "On the Sampling System with Probability Proportionate to Sum of Sizes," *Annals of the Institute of Statistical Mathematics*, 3(1), 99–107.

Parker, P. A., Janicki, R., and Holan, S. H. (2023), "A Comprehensive Overview of Unit-Level Modeling of Survey Data for Small Area Estimation under Informative Sampling," *Journal of Survey Statistics and Methodology*, DOI: 10.1093/jssam/smad020.

Parker, P. A., Holan, S. H., and Janicki, R. (2022), "Computationally Efficient Bayesian Unit-Level Models for non-Gaussian Data under Informative Sampling with Application to Estimation of Health Insurance Coverage," *The Annals of Applied Statistics*, 16(2), 887–904.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998), "Parametric Distributions of Complex Survey Data under Informative Probability Sampling," *Statistica Sinica*, 8, 1087–1114.

Pfeffermann, D., and Sverchkov, M. (2007), "Small-Area Estimation under Informative Probability Sampling of Areas and within the Selected Areas," *Journal of the American Statistical Association*, 102(480), 1427–1439.

Savitsky, T. D., and Toth, D. (2016), "Bayesian Estimation under Informative Sampling," *Electronic Journal of Statistics*, 10(1), 1677–1708.

Si, Y., Pillai, N. S., and Gelman, A. (2015), "Bayesian Nonparametric Weighted Sampling Inference," *Bayesian Analysis*, 10(3), 605–625.

Sun, A., Parker, P. A., and Holan, S. H. (2022), "Analysis of Household Pulse Survey Public-Use Microdata via Unit-Level Models for Informative Sampling," *Stats*, 5(1), 139–153.

Tillé, Y., and Matei, A. (2016), *sampling: Survey Sampling* R package version 2.8.

Vandendijck, Y., Faes, C., Kirby, R. S., Lawson, A., and Hens, N. (2016), "Model-Based Inference for Small Area Estimation with Sampling Weights," *Spatial Statistics*, 18, 455–473.

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.