A Bayesian-Frequentist Integrated Approach to Small Area Estimation

Avinash C. Singh
NORC at the University of Chicago, Chicago, IL 60603
singh-avi@norc.org

Abstract

In the context of Census Bureau's Small Area Health Insurance Estimation (SAHIE) program, we propose a nonlinear hierarchical Bayes (HB) model with log link function such that it can be cast into a log linear mixed with additive random components (LLMARC) model which makes it amenable to usual frequentist Linear Mixed model (LMM)-type treatment. The reason for this is that neither Bayesian nor frequentist solution by itself is adequate for the desired features of the final small area estimates (SAEs) such as frequentist-type built-in benchmarking of lowest level (i.e., smallest domain level) synthetic estimates to reliable higher level direct estimates, and hierarchical ratio-adjustments of lower level SAEs to higher level SAEs (or direct estimates) for internal consistency; and benchmark-adjusted posterior distributions for point, variance, and interval estimation. To this end, a hierarchically- aggregated frequentist with hierarchical Bayes (HFHB) method is proposed which uses LLMARC to build in frequentist properties and a new construct of HFHB-posterior to build in Bayesian properties in the final SAEs. The proposed HFHB solution starts with a building block model so that a single multivariate model can be used for both county and state level estimates in the SAHIE context. Modeling of estimated domain total counts instead of estimated domain means is considered because domain counts can be aggregated to derive a group level model without changing model parameters but with a stable observed variance-covariance matrix. Under LLMARC, we can obtain a closed form expression of the marginal mean after integrating out the random effect which permits model diagnostics somewhat analogous to the usual LMM case. The HFHB-posterior allows for different conditioning datasets for different parameters and updating of posteriors which is needed to preserve higher level SAEs and to use them as benchmarks for lower level SAEs. The HFHB method is also used to preserve previous time point estimates while updating current estimates through state space modeling as a means to take advantage of temporal correlations.

Keywords: Additive Random Components with Non-linear Mean Function; Building-Block Model; Built-in Benchmarking; Frequentist-type Model Diagnostics; Hierarchical Bayes; Marginal-Conditional MCMC

1. Introduction

The problem considered in this paper is complicated and challenging but very interesting and rather practical. It grew out of the desire to consider alternatives to currently used hierarchical Bayes (HB) methodology in US Census Bureau's Small Area Health Insurance Estimation (SAHIE) program based on the American Community Survey (ACS) data. SAHIE is an important and useful program for producing health insurance coverage (IC) by income-to-poverty ratio (IPR) categories for various small areas or domains defined by geo-demographics (county by age by sex by race/ethnicity). It uses a sophisticated nonlinear HB methodology combined with frequentist-type ratio adjustments to benchmark lower level small area estimates (SAEs) to higher level estimates; see Bauder, Luery, and Szelepka (2012). The complexity of the SAHIE model seems to arise naturally as it presents a difficult application of SAE with random benchmark controls for internal consistency, robustification, and face validity. Nevertheless for the sake of comparison and for establishing validity of the current methodology, it would be useful to have an alternative methodology under a weaker set of assumptions with valid frequentist-type model diagnostics for building user confidence, and promoting wide acceptance and demand of SAHIE estimates.

In this paper we propose a Bayesian-frequentist integrated approach which is motivated from the observation that neither Bayesian nor frequentist existing methods themselves provide an adequate solution for the problem under consideration; see Rao (2003) and Pfeffermann (2013) for a review of existing methods. It is somewhat in the same spirit as the calibrated Bayes approach of Little (2011) where inferences considered under a model are Bayesian but frequentist methods are used for model development and model checking. The proposed Bayesian-frequentist integrated approach, however, goes beyond calibrated Bayes in that both frequentist and Bayesian aspects play an equally important role and hence the need for an integrated approach for making inferences. It may be remarked that

in the case of linear mixed models (LMM), both frequentist and Bayesian methods do lead to similar inferences; see e.g., Singh, Stukel and Pfeffermann (1998). The proposed method for nonlinear mixed models termed hierarchically aggregated frequentist with hierarchical Bayes (HFHB) allows for several desirable features such as provision of a single lowest level or building block model for county and states, modeling estimated domain totals instead of estimated domain means, avoiding the need for smoothing of the observed variance-covariance matrix by grouping direct B-level estimates and fitting the original model at the group level which also helps in the normality assumption of the observation error, incorporating spatial and temporal correlations as surrogates for additional covariates or predictors, and variance and interval estimation of final estimates ratio-adjusted to higher level SAEs as benchmarks. The HFHB method for nonnegative outcome variables uses a log linear mixed model for the mean function with normal random effects in order to obtain a log linear mixed model with additive random components (LLMARC) model by separating out the random effect from the nonlinear mean function after integration; see Singh and Verret (2006). The LLMARC transformation allows for LMM-type frequentist diagnostics, and built-in benchmarks to higher level direct estimates. The HFHB method also uses a new construct termed HFHB-posterior to provide benchmark-adjusted posteriors of the final SAEs based on different datasets in the aggregation hierarchy of B-level direct estimates for SAE modeling at different levels along with ratio-adjustments for internal consistency.

The organization of this paper is as follows. The formulation of the SAE problem under consideration is described in Section 2 in the SAHIE context. Limitations of a pure Bayesian approach and a pure frequentist approach are discussed in Sections 3. The frequentist aspects of the proposed HFHB solution are described in Section 4 followed by the Bayesian aspects in Section 5. A stepwise description of the HFHB solution is given in Section 6 along with its generalization to two time points to allow for temporal correlations between random effects. The final Section 7 contains summary and concluding remarks.

2. Formulation of the SAE Problem

2.1 Aggregate Level Mixed Log Linear Model: Consider modeling of a nonnegative multivariate outcome vector such as counts for the two IC categories by five IPR categories (i.e., the outcome vector with dimension 10) over domains defined by county (3142 in all) and ARSH (4 age by 4 race/ethnicity by 2 sex categories); i.e., over 100,000 domains where ARSH stands for age, race, sex and Hispanic origin. These domains can be termed as building-block (B) domains since they correspond to the lowest or smallest level at which covariate information is available for at least one covariate. Besides, all domains of interest for SAHIE can be obtained as combinations of them. In practice, however, there are likely to be B-domains with no sample or very few observations. In the following, we will assume, for simplicity, only one outcome or count variable y. It is also assumed that the mean function is linear in the log scale. Now consider a B-level aggregate model for Horvitz-Thompson estimated totals $\{t_{yb}\}$ of the domain population totals $\{T_{yb}\}$, $b = 1, \ldots K_B$, given by

$$t_{yb} = T_{yb} + e_{yb} \; = N_b \left(\mu_{yb} + \bar{E}_{yb}\right) + e_{yb} \cong N_b \mu_{yb} + e_{yb} \;\; , \tag{1a}$$

where μ_{yb} is the superpopulation model mean assumed to be linear in the log scale; i.e.,

$$\log(\mu_{yb}) = \; \boldsymbol{A}'_{xb}\boldsymbol{\beta} + \eta_{yb} \;\; , \tag{1b}$$

and where $\bar{E}_{yb}$ is the finite population error which is negligible relative to the mean parameter μ_{yb}, N_b is the known population count for domain b, A_{xb} is the column vector of aggregate level averages of p-covariates (x) such as indicators of main effects of ARSH and IPR categories, and interactions between them as well as continuous covariates used in the current SAHIE model, $\boldsymbol{\beta}$ is a p-vector of regression coefficients, the B-level random effects η_{yb}'s are independent and identically distributed $N(0, \sigma_\eta^2)$, and are independent of the sampling errors $(e_{yb})_{1 \le b \le K_B} \sim WS(0, \boldsymbol{V}_B)$ where WS stands for 'wide sense' in that only first two moment assumptions are made. In the multivariate case, the $\boldsymbol{\beta}$-parameters will be different for different y-variables along with a general covariance structure of the vector $\boldsymbol{\eta_{yb}}$ with common variances and covariances over domains b. The matrix $\boldsymbol{V}_B$ is the theoretical (Horvitz-Thompson) design-based variance-covariance (V-C) matrix of the vector $(y_b)_{1 \le b \le K_B}$ at the B-level; i.e., it is the V-C matrix of observation or sampling errors. Domains of interest are at b or higher levels denoted by d; d

=1,…, K_D. Under the above model, we will denote by θ_{yd} the small area or domain total parameters of interest which are very close to the original parameters T_{yd} except for the terms involving $N_b\bar{E}_{yb}$ and are defined as

$$\theta_{yd} = \sum_{b=1}^{m_d} \theta_{yb},\ \theta_{yb} = N_b\mu_{yb} \tag{2}$$

2.2 Modeling Domain Totals instead of Domain Means: For the above problem, it is preferable to consider aggregate level models over unit level models because informative sampling designs can be easily taken into account as in Fay and Herriot (1979). In many applications, in fact, unit level modeling may not even be feasible because of lack of some covariate information at the unit level for all units in the domain or for confidentiality reasons. It may also be preferable to consider modeling of estimated domain totals instead of domain means in order to avoid ratio bias in mean specification and to avoid bias in estimating the V-C matrix through Taylor linearization because of small or zero sample size at the B-level domains. It may be remarked that for SAE, in general, a model of direct estimates of domain means is not equivalent to the corresponding model of direct estimates of domain totals (see Rao, 2003, Sections 5.2 and 10.4) because sampling weights used to obtain direct estimates may not have been post-stratified to known domain counts. For example, in model (1), dividing the equation throughout by the domain count estimator $\widehat{N}_b$ to obtain a model for the estimated domain mean does not cancel out N_b unless $\widehat{N}_b = N_b$. Usually $\widehat{N}_b$ is not equal to N_b because post-stratification or ratio-adjustment to a large number of domain population counts at the B-level is likely to impact negatively on the precision of direct estimates; i.e., instead of making direct estimates more efficient, they could make them more unstable. The main reason for this undesirable effect is that domains of interest may have very small sample size resulting in extreme weight adjustment factors. In fact, for domains with zero sample size, such post-stratification is not even feasible.

There is another important reason to prefer to work with estimated domain totals at the B-level instead of means in that domains could be grouped together to obtain a G-level model for domain totals with the same parameters but with a more stable estimated V-C matrix $\boldsymbol{V}_G$. The B-level V-C matrix $\boldsymbol{V}_B$ is conceptually well defined but its estimate may be very unstable or unreasonable because a reasonable estimate may not even be feasible due to zero sample size. We may not wish to model $\boldsymbol{V}_B$ in the interest of minimizing assumptions; a clearly desirable philosophy in practice driven primarily by frequentist considerations. We remark that the grouping of building-blocks to stabilize the V-C matrix should not be governed by smallness of realized sample size in domains but from substantive consideration such as similarity of counties with respect to IPR within states. This way, unbiasedness of direct estimates at the group level remains valid from the standard theory of domain estimation, and a stable estimate of the corresponding design-based V-C matrix $\boldsymbol{V}_G$ can be computed. The V-C matrix at the G-level can be estimated, in general, from standard design-based considerations for ACS—either under the with-replacement primary sampling unit (PSU) assumption within each stratum or using a replication method such as the successive difference replication used for ACS. Alternatively, approximations using high entropy considerations (Deville, 1999; Berger and Tillé, 2009) can also be used.

2.3 Grouping to Avoid Smoothing of the V-C Matrix and for Normality of the Observation Error: Typically in SAE the observation error V-C matrix $\boldsymbol{V}_B$ is smoothed using generalized variance functions and then treated as known. In the present approach after grouping, this may not be required for the observation error V-C matrix $\boldsymbol{V}_G$. To see this, note that for the domain g estimator t_{yg}, the Sen-Yates-Grundy form of the design-based variance estimator $var_\pi(t_{yg})$ for a random sample s of fixed size n is given by

$$var_\pi(t_{yg}) = (-\tfrac{1}{2})\ \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(\frac{y_i\delta_{i(g)}}{\pi_i}-\frac{y_j\delta_{j(g)}}{\pi_j}\right)^2, \tag{3}$$

where π_i, π_{ij} are respectively first and second order sample inclusion probabilities, and $\delta_{i(g)}$ is an indicator function for unit i if it belongs to domain g. It follows that while at the b-level, events of pairs of observations with one element not in the domain b are likely to be prevalent, they are not likely to occur at higher G-level domains obtained after suitable grouping. Thus, we expect $var_\pi(t_{yg})$ to be more stable (i.e., in the sense of small coefficient of variation due to large number of sample observations and hence t_{yg} being more precise) as we move to higher levels by grouping more and more domains. Grouping also has a side benefit in that the normality assumption for the observation error at the G-level becomes more tenable. As a special case of (3) for simple random sampling

without replacement for a sample of size n from a population of size N, $\pi_i = n/N$, and $\pi_{ij} = n(n-1)/N(N-1)$, we have

$$var_{SRS}\left(N\sum_{i=1}^{n} y_i\delta_{i(g)}/n\right) = \frac{1}{2}\left(\frac{N}{n}\right)^2 \frac{N-n}{N(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i\delta_{i(g)} - y_j\delta_{j(g)})^2$$
$$= N^2\frac{N-n}{N}\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i\delta_{i(g)} - \overline{y\delta})^2 \qquad (4)$$

where $\overline{y\delta} = \sum_{i=1}^{n} y_i\delta_{i(g)}/n$.

Under G-level modeling, SAEs for counties involved in grouping (i.e., B-level SAEs within a given G-level) can still be obtained because estimates for all random effects at the B-level are obtained under the G-level model. Specifically, after grouping, the group or G-level observation equation is derived from the B-level by a simple aggregation of the dependent variable t_{yb} as given below where m_g denotes the group size:

$$\sum_{b=1}^{m_g} t_{yb} = \sum_{b=1}^{m_g} N_b \exp(\boldsymbol{A}'_{xb}\boldsymbol{\beta} + \eta_{\boldsymbol{b}}) + \sum_{b=1}^{m_g} e_b. \qquad (5)$$

In the case of SAHIE application, there are more than 100,000 building-blocks and so even after grouping, there should be sufficient degrees of freedom for fitting model parameters. In practice, this is expected to be the case in general.

2.4 Building Block Modeling for the Feasibility of the Exchangeability Assumption: Modeling at a very low level (i.e., the building block or B-level) may be preferred for the following reasons. Currently, different HB models are used for county and state levels despite each state being an aggregation of counties. Not only does it cause internal inconsistency because county SAEs may not add up to the corresponding state SAEs, but having a single fundamental B-level model to derive all higher level models is conceptually attractive and theoretically defensible. For instance, if there is exchangeability of random effects at the B-level, then there cannot be exchangeability of random effects η_g at the G-level because $\sum_{b=1}^{m_g} N_b\eta_b$ equated to $N_g\eta_g$ implies variance of η_g must be $\sigma_\eta^2 \sum_{b=1}^{m_g} N_b^2 / N_g^2$; see Singh and Yuan (2010), and Malec and Muller (2008) for related comments. As a way out, one can treat county by ARSH domains as building-blocks and have one model for both county and state level estimates.

For the B-level model, ideally building-blocks should be similar in population counts (introducing N_b as a multiplicative factor in modeling of θ_{yb} helps toward the assumption of exchangeability of $\eta_{\boldsymbol{b}}$ by accounting for differential building block sizes) and as close to the unit level as possible for which covariate information is available. Observe that for LMM, the factor N_b arises naturally if we derive an aggregate B-level model from a unit level model for the mean (θ_{yb}/N_b) where all the covariates are at B (or higher) levels; i.e., have common values for all units within the aggregation level. Note that if a covariate is at B- or a higher aggregate level, then the corresponding common value at the unit level is obtained by dividing the covariate value by the number of units in the aggregate.

3. Limitations of Pure Bayesian and Pure Frequentist Approaches

Neither a pure Bayesian nor a pure frequentist approach is adequate for the SAE problem for the following reasons.

3.1 Pure Bayesian: Limitations

The model (1a, b) at the G-level can be embedded in an HB framework (see Section 5) with a uniform prior on $\boldsymbol{\beta}$ and a vague hyper-prior on σ_η^2 to obtain the posterior distribution using Markov Chain Monte Carlo (MCMC) denoted by

$$\left[\boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)}, (\eta_b^{(G)}) \middle| \boldsymbol{t}_y^{(G)}\right] \qquad (6)$$

where the superscript G denotes the G-level data $\boldsymbol{t}_y^{(G)}$ used for estimating parameters under HB. From (6), point, variance, and interval estimates of desired D-level domain parameters θ_{yd} can be obtained. However there are several issues.

(a) BLUP-Type Interpretation and LMM-Type Frequentist Diagnostics; For LMM, the frequentist solution of best linear unbiased prediction (BLUP) of small area parameters has a closed form expression and simple shrinkage-type interpretation which can be easily understood by users. The shrinkage interpretation is convincing to users in that if the direct estimator is relatively more reliable, then the corresponding SAE exhibits less shrinkage to the synthetic estimator (defined as the mean function when the random effect is zero). With HB solution, however, there is no analytically closed form which makes it difficult to interpret it as a shrinkage estimator although in principle any Bayesian solution has less synthetic or prior influence as the direct estimator gets more precise (Kass and Steffey, 1989). Besides having a BLUP-type interpretation, it is important to be able to perform frequentist LMM-type model diagnostics to verify assumptions directly for building user confidence.

(b) Built-in Benchmarking to Reliable Higher Level Direct Estimators: With LMM or with generalized LMM (GLMM) with random effects outside the nonlinear mean function, it is possible to have built-in benchmarking of SAEs (or just synthetic estimates) to higher level direct estimators or known population totals by enlarging the model with addition of certain auxiliary variables; see Singh (2006), Singh and Verret (2006) and Wang, Fuller, and Qu (2008). In SAHIE, random benchmark constraints on parameter estimates are national level direct estimates of insured and uninsured obtained from ACS, and nonrandom benchmarks such as the total population counts. Such benchmarking is desirable in the interest of robustification against possible model misspecification and to reduce the wellknown overshrinkage problem in SAE possibly due to the exchangeability assumption; see e.g., the triple goal estimator of Louis (1984) and the constrained Bayes estimator of Ghosh (1992). However, random benchmarks cannot inherently be incorporated within a Bayesian framework because the benchmark is based on the same data that are used as input to the likelihood. This also implies that under HB it is not theoretically defensible to benchmark each MCMC realization and treat the resulting distribution as the desired benchmark-adjusted posterior distribution. For ensuring that HB satisfies benchmark constraints, a frequentist-type ratio adjustment is typically made in practice via a second step outside the HB framework or use of a penalized loss function as in Datta, Ghosh, Steorts, and Maples (2010). However, the penalized loss function framework may not be quite satisfactory because resulting SAEs are not subject to natural constraints on parameter estimation such as non-negativity. Also, it does not provide benchmark-adjusted posterior distributions of parameters of interest. Thus, the inferential properties of the resulting estimators do not fall under either frequentist or Bayesian realms.

(c) Synthetic Influence on HB-SAEs at Higher Levels: The optimal HB-SAEs of θ_{yd}'s under the squared error loss function using the G-level data are expected to exhibit more shrinkage toward the synthetic estimators than the direct estimators of domain totals because of more relative instability (which also implies less precision) of the g-level direct estimator t_{yg} than the higher d-level direct estimators t_{yd}. As the direct estimator gets more precise at higher levels, HB is expected to exhibit less shrinkage if the model at the higher level is based on summary statistics or aggregated direct estimates derived from the B-level model. However under HB estimation using the G-level data in estimating higher level small area parameters, synthetic influence persists because θ_{yd} is estimated as a sum $\sum_{g=1}^{m_d} \theta_{yg}$. This suggests that it might be better to go for a nonoptimal estimator which uses different input datasets or summary statistics for the likelihood used in HB at different domain levels in a hierarchy of aggregates of basic t_{yb} estimates; see also Blum (2010) for use of summary statistics for approximate Bayes. However, parameters for different likelihoods are all the same as those at the B-level. This is somewhat similar to the current practice of different input datasets for state and county models except that parameters are not the same unlike the proposed hierarchically aggregated models.

(d) Internal Consistency of SAEs from Different Levels: With the current practice of different models used for different domain levels, there is the problem of internal inconsistency and hence of face validity in that HB-SAEs at lower levels may not sum to SAEs at higher levels. To remedy this, hierarchical benchmarking through frequentist-type ratio adjustments is typically used in a separate step. The end result is neither Bayesian nor frequentist. How to find variance and interval estimates that are adjusted for hierarchical benchmarking needs to be addressed.

3.2 Pure Frequentist: Limitations

The GLMM of (1) with the log link function can be easily transformed into a log linear mixed with additive random components (LLMARC) by integrating out the random effect in the mean function using log normal results; see Section 4. Consequently, the conditional mean given the random effect is sum of the marginal mean (nonlinear in fixed parameters only) and a transformed random effect with a log normal distribution. It is then possible to apply frequentist quasi-likelihood (QL of which BLUP is a special case for LMM) to obtain SAEs. However, there are several issues.

(a) Intractability of Marginal Likelihood: For estimating fixed parameters ($\boldsymbol{\beta}, \sigma_\eta^2$) of the LLMARC model, the marginal likelihood is not tractable due to very high dimensional integration required for a large number (K_B) of random effects. Therefore, standard methods such as ML (for both $\boldsymbol{\beta}, \sigma_\eta^2$) or REML (for σ_η^2) are not applicable. However, other semi-parametric frequentist methods such as quasi-likelihood under first two moment specifications or REML after linearizing the mean function under LLMARC with a working assumption of normality of error terms can be used to obtain consistent estimates as in Singh and Verret (2006).

(b) Inadmissibility of Second Order Parameter Estimates: The semi-parametric frequentist methods mentioned above are known to have problems in general for providing admissible estimates of second order parameters (variances and correlations). It is a serious drawback of such methods, and several remedies are proposed in the literature; see e.g., Wang and Fuller (2003), Li and Lahiri (2010) among others. However, these are not fully satisfactory for our purpose due to nonnormal random effects under LLMARC modeling. In such problems, HB methods are known to produce reasonable estimates (Bell, 1999).

(c) MSE of SAEs Adjusted for Estimated Second Order Parameters and Benchmarking: Even in the absence of benchmarking, it seems difficult to use available frequentist methods to avoid underestimation of mean square error (MSE) of SAEs in order to account for estimated second order parameters. The main reason is that random effects under LLMARC are not normal. The problem gets further compounded for adjusting MSE for hierarchical benchmarking to higher level direct or indirect (i.e., SAE) estimates.

(d) Interval Estimation of Small Area Parameters: Usual frequentist methods assume approximate normality to construct interval estimates from point estimates and MSE estimates. However, for LLMARC with non-normal random effects, the assumption of approximate normality of SAEs is not tenable.

4. Frequentist Aspects of the Proposed HFHB Solution

Here we assume that estimates of the fixed parameters ($\boldsymbol{\beta}, \sigma_\eta^2$) are already computed using HB at the G-level (see Section 5 on the Bayesian aspects of HFHB), and consider quasi-likelihood (QL) estimation of random effects and small area parameters under LLMARC. The LLMARC model can be derived from the GLMM of (1) because with the log link function, an analytically closed form of the marginal mean is obtained by integrating out $\boldsymbol{\eta_b}$**'s.** Specifically, using properties of log-normal distribution, the model (1b) can be recast in terms of the marginal mean as follows.

$$\theta_{yb} = N_b(e^{\boldsymbol{A}_{xb}'\boldsymbol{\beta}+\sigma_\eta^2/2} + e^{\boldsymbol{A}_{xb}'\boldsymbol{\beta}}\lambda_{yb}), \quad \log(\lambda_{yb}+e^{\sigma_\eta^2/2}) \equiv \eta_{yb} \sim_{ind} N(0,\ \sigma_\eta^2) \qquad (7)$$

where the additive random effect λ_{yb} (= $e^{\eta_{yb}} - e^{\sigma_\eta^2/2}$) has mean 0 and variance $e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$. Observe that while the random effect is transformed under LLMARC, the regression parameters $\boldsymbol{\beta}$ do not change except for the intercept. Also note that LLMARC is applicable in general because the y-variable is typically nonnegative or can be made so after translation. Also in the case of multivariate models with count data, the log link function may not be very restrictive if there are no parameter constraints such as G-level constraints of total counts (N_g) on components

of the multivariate vector ($\widehat{\boldsymbol{\theta}}_{yg}$) of count estimates. The above model provides a direct connection between a frequentist QL solution and an HB solution where the frequentist model does not require hyper-prior specification but matches the V-C specifications for the observation error and random effects under HB.

4.1 QL Estimation of Small Area Parameters: Consider the G-level version of LLMARC obtained after grouping the B-level direct estimates. Denoting $\sum_{b=1}^{m_g} t_{yb}$ by the gth group total t_{yg}, the corresponding mean $\sum_{b=1}^{m_g} \theta_{yb}$ by θ_{yg} and the corresponding sampling error $\sum_{b=1}^{m_g} e_{yb}$ by e_{yg}, we have the LLMARC at the G-level as

$$t_{yg} = \theta_{yg} + e_{yg},\ (e_{yg})_{1\le g\le K_G} \sim N(0, \boldsymbol{V}_G) \tag{8a}$$

$$\theta_{yg} = \sum_{b=1}^{m_g} N_b \nu_{yb} + e^{-\sigma_\eta^2/2} \sum_{b=1}^{m_g} N_b \nu_{yb} \lambda_{yb}, \quad \log(\lambda_{yb} + e^{\sigma_\eta^2/2}) \equiv \eta_{yb} \sim_{ind} N(0,\ \sigma_\eta^2) \tag{8b}$$

where $\nu_{yb} = e^{\boldsymbol{A}_{xb}'\boldsymbol{\beta} + \sigma_\eta^2/2}$. Given $\boldsymbol{\beta}$ and σ_η^2, a frequentist (QL) estimator of θ_{yg} using optimal estimating functions or quasi-likelihood is obtained as

$$\hat{\theta}_{yg,QL} = \sum_{b=1}^{m_g} N_b \nu_{yb} + g^{th}\ element\ of\ \boldsymbol{U}_{\mathrm{G}} \boldsymbol{W}_G^{-1} \left(t_{yg} - \sum_{b=1}^{m_g} N_b \nu_{yb}\right)_{1\le g\le K_G}, \tag{9a}$$

where the total V-C matrix $\boldsymbol{W}_{\mathrm{G}}$ is $\boldsymbol{V}_{\mathrm{G}} + \boldsymbol{U}_{\mathrm{G}}$, $\boldsymbol{U}_{\mathrm{G}} = diag\{(e^{\sigma_\eta^2} - 1)\left(\sum_{b=1}^{m_g} N_b \nu_{yb}\right)^2\}_{1\le g\le K_G}$. This is useful for computing residuals $t_{yg} - \hat{\theta}_{yg,QL}$ for frequentist model diagnostics. In vector form, the QL-estimator of $\boldsymbol{\theta}_y^{(G)} (= (\theta_{yg})_{1\le g\le K_G})$ is given by

$$\widehat{\boldsymbol{\theta}}_{y,QL}^{(G)} = \widehat{\boldsymbol{\theta}}_{y,SYN}^{(G)} + \boldsymbol{U}_{\mathrm{G}} \boldsymbol{W}_G^{-1} (\boldsymbol{t}_y^{(G)} - \widehat{\boldsymbol{\theta}}_{y,SYN}^{(G)}), \quad \widehat{\boldsymbol{\theta}}_{y,SYN}^{(G)} = \left(\sum_{b=1}^{m_g} N_b \nu_{yb}\right)_{1\le g\le K_G} \tag{9b}$$

The above estimator $\widehat{\boldsymbol{\theta}}_{y,QL}^{(G)}$ is a shrinkage estimator in that it gets closer to the direct estimator $\boldsymbol{t}_y^{(G)}$ than the synthetic estimator $\widehat{\boldsymbol{\theta}}_{y,SYN}^{(G)}$ as the direct estimator gets relatively more precise; i.e., when the number of observations in each G-level domain is large such that the shrinkage matrix gets close to the identity matrix. In other words, eigenvalues of $\boldsymbol{U}_{\mathrm{G}} \boldsymbol{W}_G^{-1}$ which are always between 0 and 1 get close to 1 or alternatively, eigenvalues of the signal to noise ratio matrix $\boldsymbol{U}_{\mathrm{G}} \boldsymbol{V}_G^{-1}$become large. It follows that under the aggregation hierarchy of direct estimators used as input datasets at different levels, QL-SAEs at higher levels are likely to be closer to direct estimates as compared to the estimates obtained by aggregating SAEs from the G-level; see Subsection 2.3 about the precision of the V-C matrix after grouping. In particular, QL-SAEs at the state level based on the revised model after aggregating to the state level are likely to exhibit less shrinkage to the state level synthetic estimates. The reason for this is that the V-C matrix at the state level is likely to be relatively more precise than the V-C matrix at the lower *G*-level which is based on unbalanced data with clustering at 0 because many data points in the state sample may not belong to any given domain at the *G*-level. However, an interesting observation was made by Bell (2013) that this may not be the case if *G*-level domains are themselves strata although such a scenario may not be practical because the *G*-level represents a fairly low level of aggregation. It is also of interest to note that with the G-level or higher level modeling, QL-estimates of all the elements of the random effect vector $(\eta_{yb})_{1\le b\le K_B}$as defined in the original B-level model are obtained even though some of the B-level domains may not have any sample data.

4.2 Frequentist-Type Model Diagnostics: Besides usual Bayesian diagnostics of posterior predictive checks and cross-validation for the HB model (Section 5) corresponding to the LLMARC model, usual frequentist diagnostics can also be conducted by standardizing the residuals $t_{yg} - \hat{\theta}_{yg,QL}$ where HB estimates of second order parameters (variance components) are plugged in. It is also useful to consider Cholesky residuals (Hausman, Ryan, and Coull, 2004) so that the residuals become approximately uncorrelated because use of common estimated regression parameters induces correlations in the residuals. Note that unlike unit-level residuals where such correlations could be ignored, it is not reasonable to do the same with aggregate or domain level residuals because domain sample sizes relative to the total sample size do not tend to zero. In addition to testing normality of standardized residuals, estimated random effects η_{yb} under LLMARC can be standardized to check for normality.

Moreover, it is important to perform both internal and external validation of the LLMARC model. As part of internal validation, besides residual diagnostics, estimates under alternate models with simplified assumptions should be examined for comparison with the chosen model. Although the scope of external validation in general is limited due to lack of alternative independent estimates, out-of-sample residuals in predicting observed domain estimates not used in modeling can serve as a surrogate for external validation. This is in the same spirit as cross-validation residuals but computationally less intensive because domains can be ordered like a pseudo-time series in the decreasing order of observed sample size, and then state space modeling can be used for one-step and two step ahead prediction residuals; see Singh (2006) for the case of LMM. Finally, it would be of considerable interest to plot $\hat{\theta}_{yg,HB}$ against $\hat{\theta}_{yg,QL}$, and $\hat{\eta}_{yb,HB}$ against $\hat{\eta}_{yb,QL}$, and check their clustering around the 45 degree line.

4.3 Built-in Benchmark Constraints to Higher Level Direct Estimates: Under GLMM, it is possible to have built-in benchmarks (random or nonrandom) whenever the marginal mean after integrating out the random effects has an analytically closed form; Singh and Verret (2006) for GLMM, and Singh (2006), and Wang, Fuller and Qu (2008) for LMM. In the case of LMM, extra covariates are introduced such that the additional estimating functions for the new fixed regression parameters are identical to benchmark constraints. The modified estimating functions for regression parameters are given by

$$\boldsymbol{X}_{+}^{(G)\prime}\boldsymbol{W}_{G}^{-1}\left(\boldsymbol{t}_{y}^{(G)}-\boldsymbol{X}_{+}^{(G)}\hat{\boldsymbol{\beta}}_{+}^{(G)}\right)=\boldsymbol{0} \tag{10}$$

where the subscript '+' in the covariate matrix $\boldsymbol{X}_{+}^{(G)}$ denotes that it is extended and the superscript 'G' denotes the data for model fitting is at the G-level. In the case of nonlinear models for the marginal mean as in LLMARC, the mean is first linearized by Taylor and then the linear estimating equation (10) is solved iteratively to satisfy benchmarks exactly. Specifically, if it is desired to benchmark SAEs to higher level direct estimates, then the extra covariates take the form $\boldsymbol{V}_G\boldsymbol{\delta}_c$ where $\boldsymbol{\delta}_c$ is a vector of 0s and 1s indicating presence or absence of a domain from the G-level in the benchmark. On the other hand, if it is desired to benchmark synthetic estimates which may be preferable to protect against misspecified models, then the extra covariates take the form $\boldsymbol{W}_G\boldsymbol{\delta}_c$. In the latter case, the new covariates involve unknown second order parameters (σ_η^2) and so these covariates are redefined iteratively. Under the HB modeling of the next section, these iterations become automatically part of MCMC replicates. In fact, even when benchmarking SAEs and not just the synthetic estimates under LLMARC, the Taylor linearized mean function also involve σ_η^2 (Singh and Verret, 2006). For the HFHB method, we propose using benchmarking of synthetic estimates only because the hierarchical benchmarking step ensures that the SAEs sum to higher level direct estimates anyway. Note that the extra covariates created for benchmarking involve the V-C matrix $\boldsymbol{V}_G$. However, if SAEs at the B-level are also of interest, the V-C matrix $\boldsymbol{V}_B$ is needed which may not be available. For this purpose, the available matrix $\boldsymbol{V}_G$ at the higher G-level can be expanded to obtain $\boldsymbol{V}_B$ in an internally consistent manner by assigning to each element in the matrix $\boldsymbol{V}_B$ in a given group the corresponding value in the matrix $\boldsymbol{V}_G$ divided by the number of domains in the group as in Singh and Verret (2006).

4.4 Hierarchical Benchmarking for Internal Consistency: Under a hierarchical aggregation of domains, higher level domain direct estimators are used as input to the model in order to reduce the shrinkage to the synthetic component of SAE. As a result, SAEs at different levels of the hierarchy may not be internally consistent. To overcome this problem of lack of internal consistency and consequently lack of face validity, the SAEs from different levels can be ratio-adjusted (or raked if there is more than one dimension) to higher level SAEs in a top-down manner. Such type of ratio-adjustment is commonly made in survey sampling for post-stratification. Specifically, for QL-SAEs, we adjust the D-level SAEs $\hat{\theta}_{yd,QL}$, d = 1, …, m_h to sum to the immediately preceding higher level SAE $\hat{\theta}_{yh,QL}$for domain h containing m_h lower D-level domains to obtain the ratio-adjusted SAEs $\tilde{\theta}_{yd,QL}$. That is, we obtain

$$\tilde{\theta}_{yd,QL}=\hat{\theta}_{yd,QL}\left(\hat{\theta}_{yh,QL}/\sum_{d=1}^{m_h}\hat{\theta}_{yd,QL}\right). \tag{11}$$

5. Bayesian Aspects of the Proposed HFHB Solution

Suppose extra covariates are included in the model (1) at the G-level so that benchmark constraints are exactly satisfied by QL-SAEs and that the model is already validated using frequentist diagnostics. Now consider an HB version of the model as given below.

Level 1: $t_{yg} = \theta_{yg} + e_{yg},\ (e_{yg})_{1\le g\le K_G} \sim N(0, \boldsymbol{V}_G))$,
Level 2: $\theta_{yg} = \sum_{b=1}^{m_g} N_b \exp(\boldsymbol{A}'_{xb}\boldsymbol{\beta} + \eta_{yb}),\quad \eta_{\boldsymbol{yb}} \sim_{iid} N(0, \sigma_\eta^2)$, (12)
Level 3: Vague hyper-prior for inverse σ_η^2 such as *Gamma* (a, b) with very small a, b > 0, and improper (uniform) prior for $\boldsymbol{\beta}$.

5.1 Admissible Estimation of σ_η^2 and Approximate Benchmarking of HB-Synthetic Estimates: Using MCMC replicates ($\boldsymbol{\beta}^{(G,k)}$, $\sigma_\eta^{2(G,k)}$), k= 1,…, M, from the posterior distribution $\left[\boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)} \middle| \boldsymbol{t}_y^{(G)}\right]$ under the HB model (12), we obtain posterior means as point estimates $\widehat{\boldsymbol{\beta}}_{HB}^{(G)}$ and $\hat{\sigma}_{\eta,HB}^{2(G)}$. The estimator $\hat{\sigma}_{\eta,HB}^{2(G)}$ is admissible (i.e., positive) by construction and hence appealing; see also Bell (1999). Moreover, since ML and HB estimates of regression parameters are asymptotically equivalent (Kass and Steffey, 1989), and since the QL estimate is consistent and close to ML if σ_η^2 is small or if the marginal errors $t_{yg} - \sum_{b=1}^{m_g} N_b \nu_{yb}$ are approximately normal, we expect $\widehat{\boldsymbol{\beta}}_{HB}^{(G)}$ to be fairly close to $\widehat{\boldsymbol{\beta}}_{QL}^{(G)}$. Therefore, it follows from subsection 4.3, the HB-SYNs are expected to satisfy approximately the benchmark constraints of reliable higher level direct estimates.

5.2 The HFHB-Posterior Construct to Account for Different Hierarchical Aggregation of Direct Estimates as Data Input at Different Levels: As noted in subsection 4.1, given fixed parameters ($\boldsymbol{\beta}, \sigma_\eta^2$), the QL estimator shows less shrinkage to synthetic and becomes closer to direct as direct estimates get more precise. This is expected to happen if hierarchical aggregation of direct estimates is used at higher levels as input to the model. Under Bayesian, since the posterior is proportional to the product of the likelihood and the prior and if the maximum likelihood estimate of the parameter of interest is relatively more precise than provided by the prior, then the HB estimate gets closer to the maximum likelihood estimator; see e.g., Kass and Steffey (1989). This implies that given $\boldsymbol{\beta}, \sigma_\eta^2$, the HB estimator is likely to exhibit less model dependence if direct estimates for the small area model at any given level are based on the corresponding hierarchical aggregates.

To incorporate the feature of approximate built-in benchmarking of HB-SYNs to high level direct estimates and the desire to make HB-SAEs closer to direct estimates at higher levels of aggregation, a new construct termed HFHB-posterior is needed as explained below. First we compute the marginal posterior $\left[\boldsymbol{\beta},\ \sigma_\eta^2 \middle| \boldsymbol{t}_y^{(G)}\right]$ of the fixed first and second order parameters from the G-level data (lowest possible for modeling) in view of the desired approximate built-in benchmarking feature of the HFHB solution. Next starting with the highest level (denote by H) which could be the Census Division or Subdivision, we compute the conditional posterior $\left[(\eta_{yb}) \middle| \boldsymbol{\beta},\ \sigma_\eta^2, \boldsymbol{t}_y^{(H)}\right]$ using the corresponding hierarchical aggregates of direct estimates in order to make HB-SAEs closer to the direct estimates. Now due to the discrepancy between conditioning datasets $\boldsymbol{t}_y^{(G)}$ and $\boldsymbol{t}_y^{(H)}$ in the marginal and conditional posteriors needed for the joint posterior of $(\boldsymbol{\beta}, \sigma_\eta^2, (\eta_{yb})_{1\le b\le K_B})$, we need to introduce a new construct of HFHB-posterior as defined below.

Joint HFHB-posterior of $(\boldsymbol{\beta}, \sigma_\eta^2, (\eta_{yb})_{1\le b\le K_B})$ at H-level = $\left[\boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)} \middle| \boldsymbol{t}_y^{(G)}\right] \times \left[(\eta_{yb}^{(H)}) \middle| \boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)}, \boldsymbol{t}_y^{(H)}\right]$, (13)

where the super-scripts G in $\boldsymbol{\beta}^{(G)}$, $\sigma_\eta^{2(G)}$ and H in $(\eta_{yb}^{(H)})$ (short for $(\eta_{yb}^{(H)})_{1\le b\le K_B}$)) denote the level of hierarchical aggregation of the dataset used for estimating the model parameters. The two components (marginal and

conditional) of the joint HFHB-posterior are of course traditional but the product is not. Nevertheless, the joint is still a legitimate probability distribution. The above HFHB-posterior can be obtained empirically by computing two linked sets of MCMC replicates—one marginal MCMC and the other conditional MCMC as follows. The marginal MCMC replicates are obtained from the empirical joint posterior $\left[\boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)}, (\eta_{yb}^{(G)})\middle|\boldsymbol{t}_y^{(G)}\right]$ using full conditionals $\left[\boldsymbol{\beta}^{(G)}\middle|\sigma_\eta^{2(G)}, (\eta_{yb}^{(G)}),\ \boldsymbol{t}_y^{(G)}\right]$ which is Normal, $\left[\sigma_\eta^{2(G)}\middle|\boldsymbol{\beta}^{(G)}, (\eta_{yb}^{(G)}), \boldsymbol{t}_y^{(G)}\right]$ which is inverse Gamma, and $\left[(\eta_{yb}^{(G)})\middle|\boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)}, \boldsymbol{t}_y^{(G)}\right]$ which requires the Metropolis-Hastings (M-H) algorithm; see Rao (2003, pp. 271). The posterior $\left[(\eta_{yb}^{(G)})\middle|\boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)}, \boldsymbol{t}_y^{(G)}\right]$ is proportional to the product of the normal prior of $(\eta_{yb}^{(G)})$ given $\sigma_\eta^{2(G)}$ and the normal likelihood $[\boldsymbol{t}_y^{(G)}|\ \boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)},\ (\eta_{yb}^{(G)})]$ as defined by equation (12). For cycle k (each cycle consists of a full set of realizations of components of $(\boldsymbol{\beta}, \sigma_\eta^2, (\eta_{yb})_{1\le b\le K_B})$), and given a replicate $(\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)})$ from $\left[\boldsymbol{\beta}^{(G)},\ \sigma_\eta^{2(G)}\middle|\boldsymbol{t}_y^{(G)}\right]$, the conditional MCMC is run to obtain a replicate $(\eta_{yb}^{(H,k)})$ from $\left[(\eta_{yb}^{(H)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(H)}\right]$ using the M-H algorithm.

The above marginal-conditional MCMC is different from the usual MCMC in that for each cycle k, given $(\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)})$, a new conditional MCMC with a separate burn-in period is required to even obtain a single replicate $(\eta_{yb}^{(H,k)})$ for the same cycle. This ensures that the MCMC replicates for $\left[(\eta_{yb}^{(H)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(H)}\right]$ are from a stationary chain because the conditioning dataset changes from $\boldsymbol{t}_y^{(G)}$ to $\boldsymbol{t}_y^{(H)}$. From the linked replicates $(\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)})$ and $(\eta_{yb}^{(H,k)})$ for each cycle k = 1,…, M, the HFHB-posterior means provide the proposed HFHB-SAEs $\hat{\theta}_{yh,HFHB}$;h= 1,…, K_H at the H-level. However, before we designate $(\eta_{yb}^{(H,k)})$ as the replicate estimate from cycle k, a number of replicate values of $(\eta_{yb}^{(H,k)})$ for the same cycle k at level H are generated so that separate burn-in periods can be completed for each level starting with the next lower level D, and further lower levels until the lowest level G is reached. These extra replicates serve as candidates for the conditional MCMC at lower levels down the aggregation hierarchy.

Finally, once the burn-in periods are completed for all hierarchical aggregation levels for cycle k, the subsequent replicates, one from each level of the hierarchy, together form the desired multivariate replicate vector $\left((\eta_{yb}^{(H,k)})', (\eta_{yb}^{(D,k)})', \dots, (\eta_{yb}^{(G,k)})'\right)'$ for cycle k, all for the same set of $\boldsymbol{\eta}$ −parameters but based on different conditioning datasets from different levels. These separate replicate estimates of $\boldsymbol{\eta}$ −parameters can be viewed as updated estimates as more and more detailed lower level datasets are used. Note that the K_B −components of the replicate $(\eta_{yb}^{(H,k)})$ from level H serve in order as M-H candidates for the K_B −components of the replicate $(\eta_{yb}^{(D,k)})$ which in turn serve as M-H candidates for the K_B −components of the next lower level replicate until we reach $(\eta_{yb}^{(G,k)})$.

5.3 Hierarchical Benchmarking of Lower Level SAEs to Higher Level SAEs: Given HFHB-SAEs for cycle k are obtained at the highest H-level in the hierarchy, HFHB-SAEs for the lower levels are computed in a top-down manner after ratio-adjustments such that they sum to the previously obtained SAEs at the higher level. To obtain the corresponding benchmark-adjusted posteriors, the HFHB methodology can be used as follows. Specifically, suppose the next intermediate level below H is D and we wish to draw a replicate consisting of K_B components from $\left[(\eta_{yb}^{(D)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(D)}\right]$ corresponding to the marginal-conditional MCMC cycle k already performed at the previous H-level. If we obtain this replicate directly from $\left[(\eta_{yb}^{(D)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(D)}\right]$ analogous to $(\eta_{yb}^{(H,k)})$ from $\left[(\eta_{yb}^{(H)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(H)}\right]$, then the replicate $(\eta_{yb}^{(D,k)})$ so obtained will not be linked to the replicate $(\eta_{yb}^{(H,k)})$ from the previous higher level although such a correspondence is, however, needed for hierarchical benchmarking.

We therefore propose to update $(\eta_{yb}^{(H,k)})$ by using more detailed domain-level information in $\boldsymbol{t}_y^{(D)}$ than $\boldsymbol{t}_y^{(H)}$. To this end, $\left[(\eta_{yb}^{(H)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(H)}\right]$ can be used as the prior for generating candidates and $[\boldsymbol{t}_y^{(D)}|\ \boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)},\ (\eta_{yb}^{(D)}), \boldsymbol{t}_y^{(H)}\ ]$ as the likelihood for obtaining the posterior $\left[(\eta_{yb}^{(D)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(D)}\right]$ using the M-H algorithm. Here as for the H-level, a separate burn-in period is required to obtain even the first replicate from this new conditional MCMC for the estimate from cycle k at level D but as explained in the previous subsection more replicates are also produced for the same cycle k to serve as candidates for the conditional MCMC at the next lower level of hierarchy. The likelihood $[\boldsymbol{t}_y^{(D)}|\ \boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)},\ (\eta_{yb}^{(D)}), \boldsymbol{t}_y^{(H)}\ ]$ can be easily obtained from the residual part of $\boldsymbol{t}_y^{(D)}$ independent of $\boldsymbol{t}_y^{(H)}$ by noting that $\boldsymbol{t}_y^{(H)}$ is a linear combination of $\boldsymbol{t}_y^{(D)}$. Specifically, consider a nonsingular linear transformation matrix (**A**) of the K_D −vector $\boldsymbol{t}_y^{(D)}$ to obtain

$$\boldsymbol{A}\boldsymbol{t}_y^{(D)} = \left(\left(t_{yh}^{(H)}\right)'_{1\le h\le K_H}, \dots, \left(t_{yd}^{(D)}\right)'_{1\le d\le m_h^{(D)}-1}, \dots\right)' \tag{14}$$

where $m_h^{(D)}$ is the number of D-level domains within the domain h at the higher H-level and $K_D = K_H + \sum_{h=1}^{K_H}(m_h^{(D)} - 1)$. Now orthogonalize the $(K_D - K_H)$ estimates $t_{yd}^{(D)}$ in $\boldsymbol{A}\boldsymbol{t}_y^{(D)}$ with respect to $\left(t_{yh}^{(H)}\right)'_{1\le h\le K_H}$ by regression to obtain the independent normally distributed data vector (denote by $\boldsymbol{t}_y^{(D|H)}$) to be used for updating $(\eta_{yb}^{(H,k)})$. The replicate values $(\eta_{yb}^{(H,k)})$ from the previous level H for the *k*th cycle serve as candidates in the M-H algorithm. The accepted candidate value $(\eta_{yb}^{(H,k)})$ or the current value gives the updated estimator $(\eta_{yb}^{(D,k)})$ and this completes the cycle *k* for the *k*th replicate from $\left[(\eta_{yb}^{(D)})\middle|\boldsymbol{\beta}^{(G,k)},\ \sigma_\eta^{2(G,k)}, \boldsymbol{t}_y^{(D)}\right]$.

For each k, the replicate SAE estimators $\{\theta_{yd}^{(D,k)}$, d =1,…, $K_D\}$ so obtained are ratio-adjusted to estimators $\{\theta_{yh}^{(H,k)}$, h=1,…, $K_H\}$ from the H-level for hierarchical benchmarking, and thus an empirical benchmark-adjusted HFHB-posterior of $\{\theta_{yd}^{(D)}$, d =1,…, $K_D\}$ is obtained. Observe that resulting posterior means $\tilde{\theta}_{yd,HFHB}$ satisfy the H-level SAEs $\hat{\theta}_{yh,HFHB}$ by construction because each replicate does. Similarly, we can obtain the HFHB-posterior at the next lower level and finally of $\{\theta_{yg}^{(G)}$, g =1,…, $K_G\}$ at the lowest G-level.

5.4 Variance and Interval Estimates Adjusted for Hierarchical Benchmarking and Estimated fixed Parameters: The point estimates at different levels of the hierarchy, although based on different datasets, satisfy benchmark constraints to hierarchical SAEs exactly. The corresponding HFHB-posteriors are automatically adjusted for hierarchical benchmarks and extra variability due to estimation of fixed parameters. Therefore, posterior variances from the HFHB-posteriors provide MSE estimates that are adjusted for extra uncertainty. Similarly, interval estimates reflect appropriate extra uncertainty. Note that with respect to the traditional posterior distribution at the lowest G-level HB model, the HFHB-estimates are nonoptimal under the squared error loss. The extra uncertainty arises because optimal estimates are not subject to hierarchical benchmarking required here for internal consistency and face validity. However, with respect to other criteria such as the frequentist MSE, the HFHB-estimates may be more efficient than the traditional HB-SAEs.

6. Proposed HFHB Solution and Its Generalization to Two Time Points

First we provide a complete stepwise description of the proposed HFHB solution whose features were described in earlier sections. We will then consider how temporal correlations can be included in modeling with data from repeated surveys such as ACS to create surrogates for additional predictor covariates and efficiency.

6.1 Description of HFHB: A stepwise description is given below.

Step I: Write the HB model at the B-level and group the B-level to the G-level for making the observation error V-C matrix stable. Now enlarge the set of covariates for approximate benchmarking of synthetic estimates from the G-level to selected high level direct estimates.

Step II: Compute the HB posterior of $(\boldsymbol{\beta}, \sigma_\eta^2, (\eta_{yb})_{1 \le b \le K_B})$ based on MCMC replicates and the M-H algorithm using the G-level data. Obtain estimates of the fixed parameters $(\boldsymbol{\beta}, \sigma_\eta^2)$ as posterior means.

Step III: Cast the HB model of Step I into an LLMARC model assuming $(\boldsymbol{\beta}, \sigma_\eta^2)$ fixed at their estimated values. Compute QL-estimates (like EBLUPs from LMM) of model parameters and perform frequentist diagnostics, both internal and external to the sample.

Step IV: Repeat Steps I to III until the model is found to be adequate.

Step V: Once the B-level model is accepted using data from the G-level, start from top at the H-level to obtain HFHB-SAEs before proceeding down to lower levels under hierarchical aggregation of B-level direct estimates. Obtain the joint HFHB-posterior from the marginal-conditional MCMC replicates ($\boldsymbol{\beta}^{(G,k)}$, $\sigma_\eta^{2(G,k)}$) and $(\eta_{yb}^{(H,k)})$ for k = 1,…, M, and the corresponding replicates $\{\theta_{yh}^{(H,k)}$, h =1,…, $K_H\}$. These replicates provide benchmarks for the next lower level.

Step VI: Given ($\boldsymbol{\beta}^{(G,k)}$, $\sigma_\eta^{2(G,k)}$) and $(\eta_{yb}^{(H,k)})$ for each replicate k from Step V, update $(\eta_{yb}^{(H,k)})$ using the next lower level (i.e., the D-level) more detailed information in terms of aggregates of direct estimates to obtain $(\eta_{yb}^{(D,k)})$ and ratio-adjust them to replicate benchmarks from the corresponding previous higher level replicate k in Step V.

Step VII: Update $(\eta_{yb}^{(D,k)})$ for each replicate k using more detailed aggregate information from other intermediate lower levels until the lowest G-level is reached. Compute hierarchically benchmarked θ −estimates at each level.

Step VIII: Compute point estimates as averages of marginal-conditional MCMC replicates for different hierarchical aggregation levels. Check for internal consistency of point estimates and their replicates. Finally obtain variance and interval estimates using replicate estimates.

6.2 Incorporating Temporal Correlation Over Two Time Points: The ACS data used for the SAHIE application are repeated over years. By combining cross-sectional and time series data, efficiency of SAEs can be considerably improved because past year estimates can provide highly correlated information for current year estimates. Here although the number of time points is small unlike usual time series applications, there are many observations (or number of domains) for each time point. A state-space model can be conveniently used to introduce dependence under both frequentist and Bayesian approaches to SAE; see e.g., Carlin, Polson, and Stoffer (1992), Singh, Mantel and Thomas (1994) and Rao (2003, pp. 162). Moreover, due to the recursive updating feature of state-space modeling via Kalman filtering, it is possible to update only current year estimates based on past observations without updating or revising previous year estimates. This provision of sub-optimal estimation is desirable in practice so that already published past year estimates are not revised, and can be conceptually justified under the proposed Bayesian-frequentsit integrated approach.

Specifically, for the HB model considered in this paper, we can assume that the regression parameters $\boldsymbol{\beta}_\tau$ do not evolve over time so that benchmarking of G-level synthetic estimates to high level direct estimates can be built-in through extra covariates at each time point. This is not likely to have much practical consequence because $\boldsymbol{\beta}_\tau$ −parameters can be estimated quite precisely from the cross-sectional data on G-level small domains; see also Singh and Roberts (1992). Besides it introduces considerable simplicity in modeling over time. A simple random

walk model can then be employed for evolution of random effects $\eta_{yb\tau}$ over time which has a reasonable implication in that η −parameters from nearby past are more correlated with the current parameter than distant past. Using G-level data from the time point $\tau - 1$, the HB-model is used to obtain the posterior of $(\boldsymbol{\beta}_{\tau-1}, \sigma^2_{\eta,\tau-1})$. Next, given a replicate value of $\sigma^2_{\eta,\tau-1}$, the posterior of $(\boldsymbol{\beta}_\tau, \sigma^2_{\eta,\tau})$ is obtained using the G-level data from τ. Here the relation $\sigma^2_{\eta,\tau} = \sigma^2_{\eta,\tau-1} + \sigma^2_\zeta$ under the random walk model is used for the posterior where σ^2_ζ denotes the variance of the random walk error term. Now starting top-down with the hierarchical aggregate data at time $\tau - 1$, we can use HFHB to compute relevant posteriors, and then use time τ data to update the posterior of σ^2_ζ from the G-level data and update $\boldsymbol{\eta}_\tau$ parameters at different levels successively from hierarchical aggregate data as in the case of a single time point. Thus HFHB posterior is used to update current year estimates without revising previous year estimates. We remark that if the sampling errors are correlated over time as is usually the case with repeated surveys, the G-level data of direct estimates at time τ need to be regressed on the time $\tau - 1$ data first to make them uncorrelated before updating current time parameter estimates using previous time point data.

7. Summary and Remarks

The problem considered in this paper is a difficult one but very interesting with a lot of practical appeal. It arose in the context of SAHIE using the ACS data. We considered an aggregate level SAE model instead of the unit level to account for informative sampling designs and a building-block model in the interest of the feasibility of the exchangeability assumption and having a single underlying model for all levels of small domains. We preferred modeling domain totals instead of area means to avoid ratio bias and the problem in approximating the observation error V-C matrix due to small or no samples. Also modeling totals allows for grouping domains for stable estimated V-C matrix and thus avoiding the problem of modeling or smoothing it. We used a log linear mixed model to satisfy range restrictions of positive mean functions.

The solution proposed arose naturally in trying to include various frequentist and Bayesian features. It required a new construct (termed HFHB-posterior) because of the use of hierarchical aggregation of direct estimated totals of lowest level domains or small areas to render SAEs close to direct estimates at higher levels and the need of internal consistency between SAEs through hierarchical benchmarking. This led to work with an HB model with a frequentist LLMARC representation for built-in benchmarking of lower level synthetic estimates. The LLMARC representation of the HB formulation allows for built-in benchmarking, LMM-type frequentist diagnostics, and a simple shrinkage-type interpretation of QL-SAEs because of an analytically closed form of SAEs. The HFHB-posterior was proposed to obtain benchmark-adjusted posteriors of HFHB-SAEs and hence their MSE and interval estimates. When dealing with two time points, the construct of HFHB-posterior can be used to update the (cross-sectional) posterior of current parameters using longitudinal correlated data without updating the (cross-sectional) posterior of previous time point parameters, and thus preserving the previously published estimates—a highly desirable feature in practice.

Besides introducing temporal correlations to take advantage of existing additional predictors via state space modeling, spatial correlations also provide additional effective predictors through simultaneous auto-regressive modeling and can be incorporated with the proposed HFHB modeling; see Rao (2003, p. 86) and Cressie (1993). In the SAHIE application, besides ACS data for direct estimates, alternative independent sources such as the 2000 decennial Census long form with income information, the IRS exemptions, and the SNAP counts for the state and county levels are also available that can be used as additional aggregate level covariates. Similarly independent estimates from Medicaid/CHIP counts for insurance coverage can be used as covariates in modeling the ACS estimates. Note that for covariates, it is not required that estimated covariate totals be unbiased for the corresponding true covariate totals. This does not lead to concerns about biased estimates of regression parameters due to random errors in these covariate totals. This is for the reason that the small area and not the original regression parameters are of direct interest and hence, we can treat covariates with measurement errors as surrogates of original covariates. The HB model of interest can be defined conditional on these surrogate covariates because their errors are independent of sampling or observation errors. Finally, we note that although the HFHB solution was presented for the univariate case, it is easily seen to be applicable to the multivariate case where the hyper-prior for the general

covariance matrix of the random effect vector can now be chosen as the flat prior or as an inverse Wishart with parameters reflecting lack of information; see Rao (2003, pp. 273). In the SAHIE context, the multivariate vector is the 10-vector of counts for IC by IPR categories as mentioned in Section 2.

Acknowledgments

This research was supported in part under US Census Bureau's Contract No. YA1323-09-CQ-0053 during the period 2012-2013. The author is grateful to Wes Basel, Don Luery, Mark Bauder and others from the Small Area Estimation Branch of US Census Bureau for discussions and useful comments. The author would also like to thank Nada Ganesh, Santanu Pramanik, and Jennifer Hasche of NORC for several useful discussions.

References

Bauder, D.M., Luery, D.M., and Szelepka, S. (2012). Small area estimation of health insurance coverage in 2008 and 2009. Technical Paper, Small Area Methods Branch, US Bureau of the Census, Washington, DC.

Bell, W.R. (2013). When does aggregation affect the amount of shrinkage in small area estimation? Technical Note, May 16, 3 pages.

Bell, W. R. (1999). Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute* 52.

Berger, Y.G. and Tillé, Y. (2009). Sampling with unequal properties. In: D. Pfeffermann and C.R. Rao, Eds., *Handbook of Statistics: Sample Surveys: Design, Method and Applications, Vol. 29A*, The Netherlands: North-Holland, 39-54.

Blum, M.G.B. (2010). Approximate Bayesian computation: A nonparametric perspective. *Jour. Amer. Statist. Assoc., Theory and Methods*, 105, 1178-1187.

Carlin, B.P., Polson, N.G., and Stoffer, D.S. (1992). A Monte Carlo approximation to nonnormal and nonlinear state-space models. *Journal of the American Statistical Association*, 87, 493-500.

Cressie, N. (1993), Statistics for Spatial Data, New York: Wiley.

Datta, G.S., Ghosh, M., Steorts, R. and Maples, J. (2010). Bayesian benchmarking with applications to small area estimation. *TEST*, 20, 574-588.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193-204.

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*, 74:269-277.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87, 533-540.

Houseman, E.A., Ryan, L.M. and Coull, B.M. (2004). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association*, 99, 383-394.

Kass, R.E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84, 717-726.

Little, R. J. A. (2011). Calibrated Bayes, for Statistics in general, and missing data in particular (with discussion). *Statistical Science*, Vol 26, No. 2, 162-174.

Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892

Louis, T.A. (1984). Estimation of a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393-398.

Malec, D. and Muller, P. (2008). A Bayesian semi-parametric model for small area estimation. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, ed. by Clarke, B.S. and Ghosal, S. Institute of Mathematical Statistics.

Pfeffermann, D. (2013), *New Important Developments in Small Area Estimation. Statistical Science*, 2013. **28**(1): p. 40-68.

Rao, J. N. K. (2003). *Small area estimation*. Wiley Series in Survey Methodology. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons].

Singh, A.C. (2006). Some problems and Proposed Solutions in developing a Small Area Estimation Product for Clients. In *Proceedings of the Survey Research Section*, American Statistical Association, pp. 3673-3683.

Singh, A.C. and Verret, F. (2006). Mixed linear nonlinear aggregate level models for small area estimation from surveys of binary counts. *Proceedings of Statistics Canada Symposium, Methodological Issues in Measuring Population Health. http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=11-522-X200600110390&lang=eng*

Singh, A.C., and Yuan, P. (2010). Building-Block BLUPs for Aggregate level small area estimation for survey data. *JSM Proceedings, Sec. Surv. Res. Meth*. http://www.amstat.org/sections/srms/Proceedings/

Singh, A.C., D.M. Stukel, and D. Pfeffermann (1998). Bayesian Versus Frequentist Measures of Error in Small Area Estimation *Journal of the Royal Statistical Society*, Series B., Vol. 60, Part 2, pp. 377-396.

Singh, A.C., Mantel, H.J. and Thomas, B.W. (1994). Time Series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33-43.

Singh, A.C. and G.R. Roberts (1992). Analysis of Cross-classified Categorical Time Series of Counts. *International Statistical Review*, Vol. 60, pp. 321-335.

Wang, J. and Fuller, W.A. (2003). The mean-squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

Wang, J., Fuller, W.A., and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 29-36.