A Bayesian-Frequentist Integration Approach to Small Area Estimation

Avi Singh

AIR-Rockville

WSS Seminar, Washington, DC

February 22, 2016

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

22 February 2016

Acknowledgments

Grateful thanks are due to Wes Basel, Bill Bell, and others in the SAE group at the Census Bureau for several useful discussions, support, and encouragement.

During 2012-13 while at NORC, preliminary ideas on BFI evolved in the process of an external review of the SAHIE program. During 2014-16, work on BFI got started and is currently underway in collaboration with Dr. Adrijo Chakraborty of NORC for potential applications to the SAIPE and SAHIE programs.

Outline

- Introduction to SAE
- Bayesian and Frequentist Approaches
- BFI in SAE for hierarchical benchmarking
- Building block small area modeling (BBSAM) of totals for compatibility between different levels of aggregation
- Grouping of BBs for stabilizing V-C matrix of sampling errors and for their approximate normality
- Modeling over time for estimating change without revising previous SAEs
- Extra covariates for built-in or self-benchmarking
- Summary

What is SAE?

- Direct estimates t_{ya} (of totals T_{ya}) are not reliable enough for lower level areas or domains but may be so for most if not all very high level areas.
- It is the sample size in the area that determines the need for SAE.
- A way out is to increase the effective sample size indirectly by modeling to connect the small area parameters θ_{ya} (i.e., by borrowing strength) where

$$T_{ya} = N_a \left(\mu_{ya} + \bar{E}_a\right) \cong N_a \mu_{ya} = \theta_{ya},\, 1 \leq a \leq K_A$$

How to model for connecting θ_{ya}?

- A Unit level Superpopulation Model:

-- LMM $y_{ak} = x'_{ak}\beta \quad + \eta_a + \varepsilon_{ak}$;

$$\varepsilon_{ak} \sim_{iid} N\,(0, \sigma_\varepsilon^2),\ \eta_a \sim_{iid} N\,(0, \sigma_\eta^2)$$

- We have $T_{ya} = N_a\left(\mu_{ya} + \bar{E}_a\right)$ where

$$\mu_{ya} = A'_{xa}\beta + \eta_{\boldsymbol{a}},\ A_{xa} = N_a^{-1}\textstyle\sum_k x_{ak},\ \bar{E}_a = N_a^{-1}\textstyle\sum_k \varepsilon_{ak}$$

- If A-level is the lowest level of availability for some x_{ak}, replace x_{ak} by A_{xa} and the unit level model is rendered into an aggregate or A-level model. Other x’s may only be available at a higher level B, then use A_{xb} to replace x_{ak}.

How to model for connecting θ_{ya}?

- The linear model has common fixed parameters β in the systematic part and σ_η^2 in the random part.
- The model could be nonlinear mixed such as log linear:

$$\mu_{ya} = e^{A'_{xa}\beta+\eta_a}$$

$$= e^{A'_{xa}\beta+\sigma_\eta^2/2} + e^{A'_{xa}\beta}\lambda_a, \quad \text{where}$$

$$\log(\lambda_a + e^{\sigma_\eta^2/2}) \equiv \eta_a \sim_{ind} N(0,\ \sigma_\eta^2)$$

- The additive random component λ_a (= $e^{\eta_{ya}} - e^{\sigma_\eta^2/2}$) has mean 0 and variance $e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$—an LLMARC model.

How do we get more efficient Est.?

- Sample randomization from the finite population:

$$\pi:\ t_{ya} = T_{ya} + e_a,\ (e_a)_{1\leq a\leq K_A} \sim (0,\ V_A)$$

- Finite population randomization from the super-population:

$$\xi: T_{ya} = N_a\left(\mu_{ya} + \bar{\bar{E}}_a\right)$$

$$\cong N_a\mu_{ya} = N_a(A'_{xa}\beta + \eta_a);\ \ \eta_a \sim_{ind} N(0,\ \sigma_\eta^2)$$

- Under joint $\pi\xi$ –randomization, we have two estimates of T_{ya}; one is the direct estimator t_{ya} and the other is the synthetic estimator $N_a\mu_{ya}$. Combine the two to obtain a more efficient composite estimator assuming V_A known.

  (Fay and Herriot, 1979).

Some General Issues of Concern

- Having constant variance of random effects at each level of aggregation but variable for different levels is not feasible because any higher level model for totals can be obtained from lower level models involving the same set of parameters. E.g., for LMM and letting B- and C- denote lower and higher levels,

$\sum_{b\in\Omega_c} N_b\,\eta_b = N_c\eta_c$ implies Var $(\eta_c) = \sigma_\eta^2 \sum_{b\in\Omega_c} N_b^2 / N_c^2$.

- The exchangeability assumption for random effects under Bayesian models at higher levels becomes questionable.
- Treating V_A as known goes against the premise of inadequate sample size for precise direct estimates. In practice, resort to generalized variance-covariance functions for smoothing.

Some General Issues of Concern

- For a given level of aggregation, areas or domains with no observations or zero contributions to the study variable are set aside during modeling and later only synthetic estimates provided as SAEs for them. This is not satisfactory because same areas when part of higher level areas play a role to obtain nonsynthetic SAEs.
- Models for sample means and totals are not equivalent unless $\widehat{N}_a = N_a$--unlikely. Modeling totals allows to have a single lowest level model from which any higher level model can be derived. Also avoids the problem of ratio bias when modeling means.

Bayesian and Frequentist Approaches

- In frequentist, inference about unknown parameters of interest is based on distributions of statistics under repeated sampling, but in Bayesian, inference is based on the posterior distribution of parameters of interest under some model for the data and prior for unknown parameters; Little (2006).

- All modeling assumptions can be validated under frequentist but not under Bayesian because of prior distribution assumptions for all parameters.

Bayesian and Frequentist Approaches

- The $\pi\xi$ –model (LMM or GLMM) introduced earlier is frequentist because no priors specified for the parameters (β, σ_η^2)—frequentist model fixed parameters. The distribution of random effects η_a, although often referred to as a prior, is just part of the model specification as it can be validated from the data (Rao and Molina , 2015, pp. 270).
- For a Bayesian $\pi\xi$ –model specification, we also need; e.g., $\beta_i \sim_{iid} N\,(0, 10^6)$, $1 \leq i \leq\ p$; $\sigma_\eta \sim U(0, 10^3)$
- Due to random σ_η^2, the random effects η_a are no longer independent unconditionally but are exchangeable.

Main Limitations of Frequentist Approaches to SAE

- **Frequentist**: Estimates of σ_η^2 may be inadmissible or unsatisfactory under usual methods ; e.g., negative (and hence truncated to 0) or may take very small values.
- For GLMM, it may be difficult in general to obtain estimates of MSE of SAEs adjusted for estimated second order fixed parameters.
- Also customary use of normality-based interval estimates not satisfactory. Some advanced methods have been developed to overcome these problems under special cases.

Main Limitations of Bayesian Approaches to SAE

- **Bayesian**: No provision of less shrinkage of higher level direct estimates to synthetic estimates when estimating higher level totals if the modeling is done at a lower level.
- If different models used at different levels, and estimates benchmarked to the higher level SAE in a second step outside the Bayesian framework, it does not provide benchmark-adjusted posteriors.
- Model diagnostics not easily understandable by users at large. In particular, interpretation of any pattern in cross-validation predictive residuals is difficult due to absence of any assumed distribution under the model.

Bayesian-Frequentist Integration for SAE

- The limitations of the frequentist can be overcome by the Bayesian and vice-versa. So BFI is a natural way to go.

- In BFI, for the Bayesian model, start with the frequentist model and then introduce priors for ‘frequentist model fixed parameters’.

BFI for SAE

- For hierarchical benchmarking, the frequentist feels free to ratio-adjust the SAEs from lower level to the corresponding SAEs at the higher level obtained from a higher level model for internal consistency and robustness to possible departures from the model. Relies on approximations for variance and interval estimation.
- The Bayesian is bound by its prescriptive rules to obtain a legitimate posterior. A serious problem arises because benchmarks based on the same data. So can't just ratio-adjust each MCMC SAE replicate.

BFI-Posterior

- Suppose the frequentist model fixed parameters (β, σ_η^2) are given and we obtain posteriors of $(\eta_b)_{1\le b\le K_B}$ from the higher level input of direct estimates $(t_{yc})_{1\le c\le K_C}$ as well as from the lower level input $(t_{yb})_{1\le b\le K_B}$ where B-level is nested within C-level. Even with a single direct estimate t_{yd} at the highest level, can get posteriors of all η_b's although their posterior means not very precise.
- We get two sets of SAEs—one at the C-level and the other at the B-level and want B-level SAEs to sum to the C-level SAEs over all b's that are nested within a given level c.
- We can obtain a benchmark-adjusted empirical joint posterior if the MCMC replicates from the two levels are linked!!!

Linking of Prior and Posterior of θ

- For any parameter θ , let $f(\theta)$ be the prior and $f(\theta|\mathrm{data})$ be the posterior. Consider two random variates θ_{prior} and θ_{post} with distributions $f(\theta)$ and $f(\theta|\mathrm{data})$ respectively. The joint distribution of θ_{prior} and θ_{post} defines the linking of the two distributions.
- The joint distribution can be obtained empirically using the Metropolis-Hastings (M-H) algorithm. For the replicate r and a U(0,1) cut-off u_r, use the candidate $\theta_{prior}^{(r)}$to obtain $\theta_{post}^{(r)}$either as the current or the candidate value depending on the acceptance probability. Thus the pairs $(\theta_{prior}^{(r)}, \theta_{post}^{(r)})$ are linked.

Linking of Lower and Higher level Posteriors of η for BFI

- For the SAE problem, given (β, σ_η^2), draw candidates $\eta_{cand}^{(r)}$ from the prior $f(\eta)$ and $u_r$from U(0,1). Now perform two separate M-H to obtain linked $\eta_L^{(r)}$ and $\eta_H^{(r)}$. This gives an empirical joint distribution $(\eta_L^{(r)}, \eta_H^{(r)})$ with *R MCMC* replicates with two datasets for the same η.
- Next obtain the MCMC replicate values of the SAE parameters from lower and higher levels and perform ratio-adjustment for hierarchical benchmarking for each replicate. The resulting empirical posterior of SAEs at the lower level is the benchmark-adjusted BFI-posterior.

How to estimate (β, σ_η^2)?

- Use hierarchical Bayes (HB) at the lowest level (the building block or B-level) for maximum efficiency.
- Obtain the posterior of all parameters $(\beta, \sigma_\eta^2, (\eta_b)_{1 \le b \le K_B})$ as the product of marginal and conditional posteriors based on different datasets for subsets of the parameters. That is, the BFI-posterior is given by

$$f(\beta, \sigma_\eta^2 | \boldsymbol{t}_y^{(B)}) \times f((\eta_b)_{1 \le b \le K_B} | \boldsymbol{t}_y^{(C)}, \beta, \sigma_\eta^2)$$

- The above BFI-posterior is nonstandard but legitimate.
- The η –parameters are estimated from the posterior conditional on the direct estimates as input at the desired level of aggregation in the interest of asymptotic design consistency (ADC).

Grouping of Building Blocks for Small Area Modeling (BBSAM)

- At the B-level, many areas may have no sample observations or may have zero contributions. The corresponding V-C matrix V_B cannot be estimated very well and the normality approximation no longer meaningful.
- Group building blocks with similar prediction scores $A'_{xb}\hat{\beta}^{(0)}$ where $\hat{\beta}^{(0)}$is obtained by fitting the regression model using nonzero t_{yb}'s at the B-level. For improving stability of the V-C matrix and normal approximation, groups can be formed such that their CVs (with synthetic estimate in the denominator) stay below a threshold.

Grouping for Building Block Small Area Modeling (BBSAM)

- The G(B)-level direct estimates are used to fit the original B-level model without changing any parameters. For HB formulation, only the level (i) specification is affected and is given by

$$t_{yg} = \sum_{b=1}^{m_g} N_b \mu_{yb} + e_g, \ (e_g)_{1 \le g \le K_{G(B)}} \sim \mathrm{N}(0, V_{G(B)})$$

- It follows that even building blocks with no observations or zero estimates take part in modeling and not kept aside. Corresponding random effects η_b's can be estimated because the group contribution in which they belong is not zero. The resulting SAEs are no longer purely synthetic.

Modeling over Time for Estimating Change without Revising Prev. Est.

- For annually repeated surveys such as ACS, beside spatial modeling, temporal modeling can be used for further efficiency gains.
- Can use SSM to model evolution of β and η $-$parameters over time with a Bayesian or a Frequentist approach.
- In estimating change $T_{ya(\tau)} - T_{ya(\tau-1)}$, using datasets from the current and previous time points yields optimal estimates. However, it is preferable not to revise already published previous estimates even though it is optimal to do so. For monthly CPS, seasonal random effects also involved; see Pfeffermann and Tiller (2006).

Modeling over Time for Estimating Change without Revising Prev. Est.

- We can use the BFI-posterior construct to estimate change without revising the previous estimate and obtain the corresponding adjusted posterior by linking MCMC replicates from $\tau - 1$ and τ.
- Recall that higher and lower level aggregates in cross-sectional modeling amounts to using coarser or less informative version of lower level input at higher levels for estimating the same set of random effects.
- Similarly, not using current time data to update previous time point estimates amounts to using less information for estimating random effects but more for current time.

Extra Covariate for Built-in or Self-Benchmarking

- For LMM, use an extra covariate $V_B 1_{b\in\Omega_h}$to obtain exact built-in benchmarking of SAEs at B-level to the direct estimator at the highest or national level under the frequentist approach; Singh (2006) and Wang, Fuller and Qu (2008). For GLMM, can define such a covariate iteratively because of unknown σ_η^2. In practice, could have several benchmarks.
- Built-in benchmarking is desirable for robustification to departures from the model; Pfeffermann (2013).

Extra Covariate for Built-in or Self-Benchmarking

- Under Bayesian, don't get exact benchmarking but inclusion of new covariates is nevertheless beneficial.
- FOR BFI, it is preferable to have built-in benchmarking of synthetic estimates rather than SAEs because lower level SAEs are benchmarked hierarchically to higher level SAEs and we want the highest level SAE to be almost identical to the direct estimator.
- For benchmarking of synthetic estimates, a different new covariate $W_B 1_{b\in\Omega_h}$is introduced where W_B=V_B+U_B, $U_B = diag\{N_b^2 \sigma_\eta^2\}$. Similarly for GLMM; Singh and Verret (2006).

Summary

- Pros and Cons for both frequentist and Bayesian approaches.
- The construct of BFI-posterior is useful to incorporate frequentist features in the Bayesian framework. Helps to alleviate the perception of the powerful Bayesian solution as a black box among practitioners often due to subjective priors and no closed form analytic estimates involving intensive Monte Carlo computations.
- Traditional Bayesian formulation cannot handle random parameter constraints based on the same data needed for hierarchical benchmarking.

Summary

- BFI-posterior was introduced to allow for different conditioning datasets for the same set of random effect parameters ; and for different subsets of parameters—frequentist model fixed and random effect parameters.
- MCMC based on M-H was used to link posteriors of random effects from higher and lower level input data or from less informative (up to the previous time point) and more informative (up to the current time point) data.

Summary

- BBSAM was used to avoid incompatibility of models of totals from different levels and for plausibility of the exchangeability assumption.
- Grouping of BBs needed to make the sampling error V-C matrix more stable, normality assumption feasible, and to obtain nonsynthetic SAEs for domains with no observations or zero direct estimates.
- Extra covariates for built-in or self benchmarking in conjunction with hierarchical benchmarking introduced for robustification to model departures.

References (Incomplete)

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

Bell, Basel, Cruse, Dalzell, Maples, O'Hara, and Powers (2007). SAIPE report, Bureau of the Census, December.

Little, R. J. A. (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap. *The American Statistician*, Vol. 60, No. 2, 1-11.

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

Pfeffermann, D., & Tiller, R. B. (2006). Small-area estimation with state–space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101(476), 1387–1397

Rao, J. N. K. and Molina, I. (2015). *Small area estimation, 2nd ed.*. Wiley Series in Survey Methodology. Hoboken, NJ: Wiley-Interscience

References (Incomplete)

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

Singh A. C. (2013). A Bayesian-frequentist integrated approach to small area. *Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference*. https://fcsm.sites.usa.gov/files/2014/05/B2_Singh_2013FCSM.pdf

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 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/

THANK YOU

Avi Singh
301-592-3349
asingh@air.org

6003 Executive Blvd, Suite 3000
Rockville, MD 20852
General Information: 202-403-5000
TTY: 887-334-3499
www.air.org