A Hierarchical Bayesian Approach to Estimation for the Annual Survey of Local Government Finances

Noah Bassel
Noah.Bassel@census.gov
Bac Tran
Bac.Tran@census.gov
U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20233

Abstract

The Annual Survey of Local Government Finances (ALFIN) is conducted by the U.S. Census Bureau and provides statistics about the financial activities of state and local governments across the nation. The Economic Statistical Methods Division (ESMD) currently uses a combination of Empirical Best Linear Unbiased Predictor (EBLUP), and Horvitz-Thompson (HT) methods to estimate these statistics. In this paper we explore a linear mixed model Hierarchical Bayes estimator. All three estimators are then evaluated through a Monte Carlo simulation experiment using data from the 2007 and 2012 Census of Governments. The performance of the three estimators is compared through their mean squared errors and relative bias.

Key words: Annual Survey of State and Local Government Finances; EBLUP; Hierarchical Bayes; Small Area Estimation

1. Introduction

Every five years, the Economic Directorate of the U.S. Census Bureau conducts a census of over 90,000 local government units to collect data on their financial activities. In the years between two consecutive censuses (years ending with 2 and 7, e.g. 2007, 2012, and 2017) the Economic Directorate also conducts the Annual Survey of Local Government Finances (ALFIN), a nationwide sample survey covering all local governments in the United States. Estimates published from the ALFIN are aggregated from the five local government types: counties, municipalities, townships, special districts, and school districts, in conjunction with data collected from the Annual Survey of School Finances. The Economic Directorate publishes local level aggregates from the ALFIN along with corresponding state level aggregates from the Annual Survey of State Government Finances. Statistics from these two surveys are used to estimate the government component of the Gross Domestic Product, allocate some federal grant funds, and provide information to assist in public policy research. More information about the ALFIN can be found at: http://www.census.gov/govs/local.

_____

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

We used two estimation methods for the 2014 ALFIN survey cycle: for each estimation cell a choice between a Horvitz-Thompson (HT) and Empirical Best Linear Unbiased Predictor (EBLUP) estimator was made based on each estimator's performance in that cell in simulation research. In prior years we had used calibration, but calibration was found to perform poorly upon further evaluation and was subsequently abandoned. While EBLUP has offered a huge improvement over HT and calibration estimators, there is room for further investigation. Additionally, there are compelling reasons to wish to conduct estimation under a fully Bayesian framework. For this research, we conducted an evaluation to compare the performance of three estimators: HT, EBLUP, and Hierarchical Bayes (HB). We used data from the 2007 and 2012 Censuses of Governments: Annual Survey of Local Government Finances (CoG-F) to carry out the evaluation.

2. Data

Data collected for the ALFIN are provided by state and local governments across the country. Each financial activity reported by local governments is assigned to an item code. These item codes can be grouped into one of four main categories: revenues, expenditures, assets and debts. Approximately four hundred item codes are included in these four categories. In the production environment, we currently select between EBLUP and HT estimators for only the expenditures and revenues item codes. For all other item codes, the HT estimator is used. The ALFIN consists of a sample of local governments along with school district data provided by the Annual Survey of State Government Finances. The annual statistics from the ALFIN data are published in two products: the downloadable file and viewable file. The downloadable file provides estimates of the total for each item code, both by state and for the nation at three different levels: local governments, state governments and combined state and local governments. In contrast, the viewable file provides aggregates of item code totals for the four main categories, as well as totals for some of the more notable detailed items. Statistics from the viewable file are given both by state and for the nation and published online in a nested table format. The coefficient of variation (CV) is provided for each estimate.

The scale of the ALFIN statistics presents formidable challenges when making estimates. During non-census years, over 30,000 state-item code totals must be estimated for the annual downloadable file. The cell sizes, based on the number of local governments contributing to the state item code estimates, are often small ($n<10$), and design-based estimators such as Horvitz-Thompson can become unstable in these conditions. This research continues earlier efforts by Schilling *et al* (2016), and Love *et al* (2013, 2014) to find alternate estimators that improve estimation stability and precision for the ALFIN.

Small area estimation (SAE), can be used to calculate estimates and measures of variability for areas, or domains, with sample sizes that are too small for direct estimation with traditional estimators such as HT. Using small area methods, the effective sample sizes can be increased by "borrowing strength" from similar domains with models and auxiliary data. Though the models can take a variety of forms, the overall goal is an appreciable increase in estimation accuracy over that of the direct estimator. Small area methods offer promise as an alternative approach to handle the challenges posed by ALFIN estimation. Auxiliary data from the CoG-F can be leveraged through models and small area methods to improve ALFIN estimates. The use of small area estimation for ALFIN is appropriate because cell sample sizes by domain (state by item code pairs) cannot be controlled and are often too small for reliable direct estimation. The small cell sample sizes are the result of a sample design that is not a direct-element design. Instead, the sampled units are local governments,

which have different combinations of item codes. The item codes associated with a local government can vary over time, and obscure item codes can be associated with small local governments, which can have low selection probabilities.

3. Sample Design

The ALFIN uses a two-phase sample design. In the first phase, a group of local governments is designated as certainties (weight=1) and included in the sample, while other local governments are selected using a stratified probability proportional-to-size (πPS) design (Särndal *et al*, 1992). In the second phase, a modified version of cutoff sampling (Dalenius & Hodges, 1959) is used to reduce the number of non-contributory municipalities, townships and special districts in the sample. This sample design was implemented in 2014 and allows the Economic Directorate to reduce sample size and respondent burden for small cities, townships and special district governments, while maintaining estimate precision and data quality. Data from the 2012 CoG-F provides the auxiliary information used for the size variable and to identify certainty units on the frame.

The sample design was implemented using a multi-step process. First, large governments were designated as initial certainty units. Next, remaining units were stratified by state and government type. Four of the five local government types (counties, municipalities, townships and special districts) were sampled by this design. Next, in the first stage of the design, a stratified πPS sample was selected, where the size variable was defined as the maximum of total expenditures and a second variable that could be total taxes, total revenues, or long-term debt, depending on the government type. Next, a cut-off point was calculated for the second stage of the design using the cumulative square root of the frequency method (Dalenius & Hodges, 1959), to distinguish between small and large government units in the municipal and special district strata. Finally, the strata with small-size government units were subsampled. For municipal strata, subsampling was carried out using a simple random sampling design; for special district strata, subsampling was accomplished through systematic sampling.

4. Estimation Methods

4.1 Parameters of Interest

In all cases we wish to obtain an estimator for the total of item code c in state k:

$$t_{kc} = \sum_{i \in U, i \in k} y_{ikc}$$

Note that one government unit (i) can have multiple item codes (c).

4.2 Direct Estimator (Horvitz Thompson)

The traditional design-based Horvitz-Thompson (HT) estimator is used as a baseline estimate:

$$\hat{t}_{kc}^{HT} = \sum_{i \in S, i \in k} w_{ik} y_{ikc}$$

Where the design weight $w_{ik} = \frac{1}{\pi_{ik}}$ and π_{ik} is the inclusion probability for unit i in state k, and sample units are summed for state k.

The HT estimator is unbiased with respect to the sample design, but can also exhibit high variance in the presence of small sample sizes such as are often present in estimation for ALFIN.

4.3 EBLUP Estimator

Consider a Linear Mixed Model (LMM) for sample data **y** on auxiliary variable (**X**) that includes both fixed (**β**) and random (**γ**) components, as shown below:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$$

This gives us the following nested unit-level model (Rao, 2003):

$$y_{ikc} = \beta_0 + \beta_1 X_{ikc} + v_{kc} + \varepsilon_{ikc}$$

Where the y_{ikc} denotes the c[th] item code value for the i[th] government unit in state k for the current year; X_{kic} a corresponding item code value from the most recent Census of Governments; β_0 and β_1 are the fixed effects (the unknown intercept and slope respectively). The v_{ic} are the random complement to the fixed X, or the small area specific random effects for our data; the ε_{kic} are errors for the individual observations $i = 1, \ldots, N_{ic}$. The distribution of the random effects corresponds to the deviations of item code log values from the value of $\beta_0 + \beta_1 X_{ikc}$. In addition, we assume that $v_{kc} \sim N(0, \tau^2)$, and $\varepsilon_{ikc} \sim N(0, \sigma^2)$. To account for the skewed nature of the data and reduce heteroscedasticity we transform the data to a log scale:

$$log(y_{ikc}) = \beta_0 + \beta_1 log(X_{ikc}) + v_{kc} + \varepsilon_{ikc}$$

Once the model has been fit and diagnostics are used to assess goodness of fit, the following model-based predictor is used for out-of-sample units:

$$\hat{y}_{ikc} = exp\left(\hat{\beta}_0 + \hat{\beta}_1 log(X_{ikc}) + \hat{v}_{kc}\right)$$

where the estimated fixed and random parameters are estimated via restricted maximum likelihood using the SAS® PROC MIXED procedure.

An estimate of t_{kc} is given by:

$$\hat{t}_{kc}^{EB} = \sum_{i \in S, i \in k} y_{ikc} + \sum_{i \in S^c, i \in k} \hat{y}_{ikc}$$

Note that $\hat{y}_{kic}$ is the model dependent predictor of the non-sampled part (S^c) of the population (U).

Because the EBLUP model borrows strength across domains (in this case across item codes) and from auxiliary information (namely the last census of governments) while allowing for domain specific variation it can significantly reduce estimator mean squared error compared to the HT, particularly in estimation cells with small sample sizes.

However, the model is dependent on normality assumptions and hence potentially sensitive to outliers.

4.4 Hierarchical Bayes

Consider now a similar nested unit-level model, but under a hierarchical Bayesian model specification which is described as follows. As before we transform to the log scale:

$$log(y_{ikc})|\boldsymbol{\beta}, \sigma_k^2, v_{kc}, \tau_k^2, X_{ikc} \sim t_4\left(\beta_0 + \beta_1 log(X_{ikc}) + v_{kc}, \sigma_k^2\right)$$
$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 1\end{bmatrix}, \begin{bmatrix}100 & 0\\ 0 & 100\end{bmatrix}\right)$$
$$v_{kc} \mid \tau_k^2 \overset{iid}{\sim} N\left(0, \tau_k^2\right)$$
$$\tau_k^2 \sim inv - \Gamma(0.01\ ,0.01)$$
$$\sigma_k^2 \sim inv - \Gamma(0.01, 0.01)$$

The choice of $\nu = 4$ degrees of freedom in the likelihood is due to Lange, Little, and Taylor (1989).

Posterior simulation is run using SAS® PROC MCMC for a sufficient number of iterations such that we have D independent samples of the parameters from their joint posterior distribution $p\left(\boldsymbol{\beta}, \tau_k^2, \sigma_k^2, v_k|\boldsymbol{y}\right)$ once burn-in and thinning are accounted for. For $d = 1, \dots, D$ using the marginal posterior samples of $\left\{\boldsymbol{\beta}^{(d)}, \boldsymbol{v}^{(\boldsymbol{d})}|\boldsymbol{y}\right\}$ we then have the following model-based predictor for out-of-sample units:

$$\hat{y}_{ikc}^{(d)} = exp\left(\hat{\beta}_0^{(d)} + \hat{\beta}_1^{(d)} log(X_{ikc}) + \hat{v}_{kc}^{(d)}\right)$$

And in turn for $d = 1, \dots, D$ we have the estimator (as before $\hat{y}_{ikc}^{(d)}$ is the model dependent predictor of the non-sampled part (S^c) of the population (U)):

$$\varphi_{kc}^{(d)} = \sum_{i \in S, i \in k} y_{ikc} + \sum_{i \in S^c, i \in k} \hat{y}_{ikc}^{(d)}$$

This gives us the HB estimator for t_{kc}:

$$\hat{t}_{kc}^{HB} = \frac{1}{D}\sum_{d=1}^{D} \varphi_{kc}^{(d)} = \varphi_{kc}^{(.)}$$

And the HB variance estimator:

$$\hat{V}(t_{kc}|\boldsymbol{y}) = \frac{1}{D-1}\sum_{d=1}^{D}\left(\varphi_{kc}^{(d)} - \varphi_{kc}^{(.)}\right)^2$$

Like the EBLUP model, the hierarchical Bayes model used here is able to borrow strength both across time and across domains, leading to improved performance over design-based estimators.

Bayesian modeling has multiple advantages over classical models that provide the motivation for its use. The hierarchical Bayesian approach to inference is straightforward, can be used to fit complicated models, and has important applications to small area

estimation. Hierarchical modeling explicitly accounts for area-to-area variation and takes advantage of the multilevel structure of the sample data. Additionally the ease with which robust modelling assumptions can be incorporated into a Bayesian framework make the hierarchical Bayesian approach more flexible than the classical EBLUP (which as mentioned previously is dependent on normality assumptions).

For this evaluation a total of 15000 iterations were run, discarding the first 2500 as burn-in. The remaining iterations of the chain were then thinned by taking every 5th observation to give D=2500.

5. Evaluation Design

This evaluation uses data from the Finance components of the 2007 and 2012 Census of Governments. The universe is the intersection of 2007 data with 2012 data, including only the units surveyed during both census years. For simplicity, the universe is further restricted to include non-zero values on the variables of interest, or the four main groups of item codes. The universe for this evaluation is comprised of approximately 85,850 units.

The 2007 CoG-F provides the auxiliary data, and serves as the sampling frame. The production sampling design is applied to select 1000 replicated samples from the 2012 CoG-F data. For each sample replicate we estimate the 2012 state totals for both expenditure and revenue item codes using the three estimators: HT, EBLUP, and HB. During the analysis, we computed the relative root mean squared error (RRMSE) and relative bias for each estimator from the 1000 samples. We performed this analysis on a subsample of 7 states (California, Idaho, New Mexico, New York, Ohio, South Carolina, and Wyoming).

5.1 Relative Root Mean Square Error (MSE)

We used the mean square error (MSE) as a primary measure for evaluating estimator quality. In this evaluation, we calculate MSE for all three estimators over all sample replicates. The MSE for a state-item code combination is calculated as $\widehat{MSE}(\hat{t}_{kc}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{t}_{kc}^{(r)} - t_{kc}\right)$, where $\hat{t}_{kc}^{(r)}$is the estimated state total of an item code for one sample replicate (r), and as before t_{kc} is the true state total of an item code. In our paper we take R= 1,000. In order to compare estimators we normalize the MSE by its corresponding cell value, giving us the Relative Root Mean Square Error (RRMSE):

$$\widehat{RRMSE} = \frac{\sqrt{\widehat{MSE}(\hat{t}_{kc})}}{t_{kc}}$$

5.2 Relative Bias

The bias of an estimator is measured as the difference between its expected value and the true value of the parameter being estimated. In our evaluation, relative bias is calculated for a state-item code combination as:

$$\widehat{RB}(\hat{t}_{kc}) = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{t}_{kc}^{(r)} - t_{kc}}{t_{kc}}$$

As noted above R=1,000 in our evaluation.

6. Results and Evaluation

A comparison of the estimated RRMSE of the three estimators is given in Table 1. The values indicate the number of times an estimator outperforms the others for $\widehat{RRMSE}$. These results are compiled from 1000 replicates with each replicate yielding state-item code estimates over 7 states for a total of 1058 possible estimates per replicate, which reduces to 836 estimation cells per replicate once we exclude state-item code estimates from cells containing only certainty units.

Table 1: Number of Times an Estimator Outperforms the Others for RRMSE (836 cells = state by item code estimates)

HT	EBLUP	HB
49	260	527

NOTE: Ties are not listed in Table 1; these can be attributed to 222 state - item code estimates that are from cells having only certainty units, with 2 exceptions, where the Calibration estimator defaulted to HT.
Data Source: U.S. Census Bureau, 2007 and 2012 Census of Governments: Finance

Table 1 demonstrates the clear superiority of small-area models over the traditional design-based HT estimator. HT is outperformed by either EBLUP or HB in approximately 94% of the estimation cells. In Table 2 we then compare only the two model-based estimators to each other.

Table 2: Direct Comparison of EBLUP and Hierarchical Bayes for RRMSE (836 cells = state by item code estimates)

EBLUP	HB
287	549

NOTE: Ties are not listed in Table 2; these can be attributed to 222 state - item code estimates that are from cells having only certainty units, with 2 exceptions, where the Calibration estimator defaulted to HT.
Data Source: U.S. Census Bureau, 2007 and 2012 CoG-F

In a direct comparison HB outperforms EBLUP approximately 65% of the time with respect to RRMSE.

Similarly a comparison of the estimated relative bias ($\widehat{RB}$) for the three estimators is given in Table 3.

Table 3: Number of Times an Estimator Outperforms the Others for Relative Bias (836 cells = state by item code estimates)

HT	EBLUP	HB
694	60	82

NOTE: Ties are not listed in Table 3; these can be attributed to 222 state - item code estimates that are from cells having only certainty units, with 2 exceptions, where the Calibration estimator defaulted to HT.
Data Source: U.S. Census Bureau, 2007 and 2012 CoG-F.

As shown in table 3 show HT outperforms the other two estimators with respect to bias over 80% of the time. This result is unsurprising because HT is unbiased with respect to the sample design.

In table 4 we directly compare the two model-based estimators to each other:

Table 4: Direct Comparison of EBLUP and Hierarchical Bayes for Relative Bias (836 cells = state by item code estimates)

EBLUP	HB
346	490

NOTE: Ties are not listed in Table 4; these can be attributed to 222 state - item code estimates that are from cells having only certainty units, with 2 exceptions, where the Calibration estimator defaulted to HT.
Data Source: U.S. Census Bureau, 2007 and 2012 CoG-F

HB outperforms EBLUP with respect to bias in approximately 59% of estimation cells. This represents an improvement, but not an enormous one.

Tables 5 and 6 provide an overall comparison of the mean $\widehat{RRMSE}$ and average relative bias for the three estimators relative to cell size. Categories are formed for median cell sizes calculated over the 1000 sample replicates for the non-certainty (πPS) units only. The last two categories show that the median cell size can be zero, indicating some state-item code combinations are obscure and have only one or two contributing πPS units, but not for every sample replicate. Two separate categories are formed for the obscure state-item code estimates, reflecting that some of these estimates can also include contributing certainty units, while others are reliant on only the πPS units.

Table 5: Overall Estimator Comparison for Mean RRMSE by Cell Size (836 cells = state by item code estimates)

Median Cell Size (πPS units only)	Number of Cells	Mean RRMSE		
		HT	EBLUP	HB
All cell sizes	836	248%	34.6%	35.2%
>30	135	6.60%	3.82%	2.78%
21-30	69	11.5%	4.44%	4.82%
11-20	108	11.1%	3.67%	2.80%
6-10	111	22.8%	6.78%	5.84%
1-5	275	42.3%	14.0%	17.5%
0*	102	67.9%	13.7%	11.9%
0**	36	5083%	603%	603%

* Includes other contributing certainty units in the estimates.
** Estimates calculated only from πPS units (no certainty units).
NOTE: Table 5 excludes 222 state-item code estimates that are from cells having only certainty units.
Data Source: U.S. Census Bureau, 2007 and 2012 CoG-F

The results from Table 5 expand on the findings from Tables 1-2, in that the two model-based estimators clearly outperform HT with regards to RRMSE over all size categories. As expected the HT estimator can still perform moderately well in cells with large sample sizes, while the superiority of the model-based approaches is clearest in cells with very small samples. Additionally while HB may offer some marginal improvements over EBLUP, its dominance is not nearly as clear as that of model-based approaches over the traditional HT. While HB may outperform EBLUP in more cells, EBLUP has a slightly lower average mean RRMSE overall, indicating that HB may have an extremely high mean-squared error in a few cells.

Table 6: Overall Estimator Comparison for Average Relative Bias by Cell Sizes (836 cells = state by item code estimates)

Median Cell Size (πPS units only)	Number of Cells	Average Relative Bias		
		HT	EBLUP	HB
All cell sizes	836	210%	23.6%	26.3%
>30	135	-0.04%	-2.37%	0.28%
21-30	69	-0.01%	-3.14%	0.56%
11-20	108	0.15%	-1.85%	-0.25%
6-10	111	-0.05%	-3.07%	-2.03%
1-5	275	1.39%	0.64%	4.55%
0*	102	-0.45%	3.91%	3.74%
0**	36	4870%	561%	570%

* Includes other contributing certainty units in the estimates.
** Estimates calculated only from πPS units (no certainty units).
NOTE: Table 6 excludes 222 state-item code estimates that are from cells having only certainty units.
Data Source: U.S. Census Bureau, 2007 and 2012 CoG-F

Similarly, the results from Table 6 expand on the findings from Tables 3-4, with HT showing by far the best performance in terms of average relative bias, whereas the two model-based estimators are very similar to each other in terms of performance. The one exception is the second obscure state-item code category, where the estimates are calculated using only πPS units. Under these conditions, the average relative bias for HT reach extreme values, while the two models outperform HT due to the presence of non-sampled units in each model. As with RRMSE, HB may outperform EBLUP in a larger number of estimation cells, but EBLUP shows lower bias on average, indicating that HB may suffer from very high bias in a few cells.

7. Conclusions

Our evaluation shows a clearly superior performance by model-based small area approaches (EBLUP and HB) with respect to $\widehat{RRMSE}$, while HT outperforms the other two estimators with respect to average relative bias due to the HT estimator's unbiasedness property. While in most cases the model-based approach is superior to HT, the Hierarchical Bayes approach offers only incremental improvements over the EBLUP mode. In particular, while the hierarchical Bayesian model offers lower $\widehat{RRMSE}$ and bias in a larger number of cells than does EBLUP, the model often underperforms EBLUP in some of the smallest cell size categories, leading to an overall higher average $\widehat{RRMSE}$ and bias. However, Bayesian models still offer certain inherent advantages over frequentist models, and the hierarchical Bayesian model introduced here can serve as an entry point for more complicated models that could potentially offer further improved performance relative to EBLUP.

8. References

Dalenius, T. and Hodges, J.L. (1959). "Minimum Variance Stratification," *Journal of the American Statistical Association,* 54, 88-101.

Lange, K.L., Little, R.J.A., and Taylor, J.M.G. (1989). "Robust Statistical Modelling Using the T-Distribution," *Journal of the American Statistical Association,* 84, 881-896.

Love, E., Barth, J. and Tran, B. (2014). "Evaluating Calibration Estimators for the Annual Survey of Local Government Finances," 2014 Joint Statistical Meetings.

Love, E. and Tran, B. (2013). "Evaluation Study of Calibration Estimation for the Annual Survey of Local Government Finance," 2013 Federal Committee on Statistical Methodology Research Conference.

Rao, J.N.K. (2003). *Small Area Estimation*, New-York, John Wiley & Sons, Inc.

Saerndal, C.E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York, Springer-Verlag.

Schilling, P., Betrouni, R., and Tran, B. (2016). "The Performance of the Empirical Best Linear Unbiased Predictor in Annual Survey of Local Government Finances," 2016 Joint Statistical Meetings.