Robust Estimation of Employment and Finance Data Using Bayesian Inference for a t-Mixture of Linear Mixed Models

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

Abstract

We present a hierarchical Bayes approach to small area estimation (SAE) for the Annual Survey of Public Employment & Payroll (ASPEP) and the Annual Survey of Local Government Finances (ALFIN). This study provides a robust estimation methodology for the total number of full-time employees in the ASPEP and the total expenditures and total revenues in the ALFIN. The estimator is based on a Linear Mixed-Effect Model (LMM) in which errors follow a mixture of $t-$distributions. We compare our research method to the existing methods being used for these surveys at the U.S. Census Bureau. The two Census of Governments (CoG) surveys for Employment and Finance for 2007 and 2012, similar to ASPEP and ALFIN for non census years, were used for the evaluation of this research.

Key Words: Linear mixed-effect models, Normal-Mixture models, $t-$Mixture models, Bayesian method, MCMC procedure, Small Area Estimation.

1 Introduction

The U.S. Census Bureau uses a hybrid approach (combination of estimators) to small area estimation for the Annual Survey of Public Employment & Payroll (ASPEP) and the Annual Survey of Local Government Finances (ALFIN). In this approach, three estimation methods are considered for each state-by-government-function estimation cell: Horvitz-Thompson (HT), Empirical Best Linear Unbiased Prediction (EBLUP), and hierarchical Bayes (HB) estimator with error terms following a $t-$distribution (HB-T). The EBLUP estimator is based on a linear mixed-effect model with errors that are assumed to be normally distributed. The HB-T estimator is based on a linear mixed-effect model with $t-$distributed errors. Linear mixed-effect models such as the ones used for EBLUP and HB-T can be sensitive to outliers. To accommodate potential outliers in the ASPEP, Trinh and Tran (2017) [14] produced a robust HB estimator with error terms following a mixture of normal distributions (HB-NN). In this research, we explore HB estimator with error terms following a mixture of $t-$distributions (HB-TT) to determine if it would perform better for these surveys.

Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information. Approval ID: CBDRB-FY19-ESMD-B00013

Census of Governments

Every five years, in years ending in “2” and “7” the Census Bureau conducts a Census of Governments (CoG), which includes both an employment component (CoG-E) and a finance component (CoG-F). The CoG-E collects exactly the same government employment and payroll data as the ASPEP collects on a sample basis in non-census years. The CoG-F also collects the same data on financial activities of all local governments as the ALFIN does on a sample basis in non-census years. Therefore, to test how the new estimator performs for both the ASPEP and ALFIN surveys, we used the California 2007 and 2012 CoG-E and CoG-F data in this research.

Annual Survey of Public Employment & Payroll and Annual Survey of Local Government Finances

The ASPEP and ALFIN surveys are designed to collect data and produce estimates of statistics on the number of state and local government civilian employees, gross payroll and expenditures, revenues, debts and assets. The target population of approximately 90,000 governmental units includes five types: counties, cities, townships, special districts, and school districts.

Each survey consists of two components. The first component contains the fifty state governments and the District of Columbia (included with certainty). The second component in the ASPEP is a representative sample of about 10,000 local governments of types 1-5. The second component in the ALFIN collects approximately 24,000 local governments containing 10,000 of types 1-4 (non-school) and 14,000 of type 5 (school district). The most recent CoG-E and CoG-F serve as the sampling frame for the ASPEP and ALFIN respectively. About two years after every CoG-E and CoG-F, the Census Bureau redesigns and selects a new sample of local governments. More information about the ASPEP and ALFIN data can be found at http://census.gov/programs-surveys/apes.html and http://census.gov/programs-surveys/gov-finances.html respectively.

2 Sample Design

Both ASPEP and ALFIN have a two-phase sample design. In the first phase, a group of selected governments is assigned as certainties (having a weight of 1, determined by subject matters) and included in the sample, while other governments are selected using a stratified systematic probability proportional to size design. In the second phase, cutoff sampling is used to reduce cost. Sample units are grouped into two strata depending on their sizes and then a subsample is selected from the stratum with small units. The details on the description of the sample design for the ASPEP and ALFIN can be found in Dumbacher and Hill (2014) [3], Bassel and Tran (2017) [2].

3 Estimators

This section briefly describes the different estimators used in this research for estimating the ASPEP total full-time employment and the ALFIN derived values for total revenues.

Let y_{mk} denote the value of the variable of interest (full-time employment or the revenues) for the k^{th} unit within the m^{th} government function (for example: hospitals, education, transportation, ...). We are interested in estimating the total $Y_m = \sum_{k=1}^{N_m} y_{mk}$ for $m = 1, ..., M$ (N_m: number of units of the m^{th} area; M: number of areas).

3.1 Direct Estimator (Horvitz-Thompson)

A direct estimator of Y_m is given by $\widehat{Y}_m^{\text{HT}} = \sum_{k=1}^{n_m} w_{mk} y_{mk}$ where $w_{mk} = \dfrac{1}{\pi_{mk}}$ and π_{mk} is the inclusion probability for the k^{th} unit within the m^{th} area.

3.2 Model-Based Small Area Estimator

A model-based small area estimator can be used to produce estimates for areas with sample sizes that are too small for direct estimation using direct estimators such as HT. Using SAE methods, the effective sample sizes can be increased by taking extra information from auxiliary data. A model-based estimator of Y_m, the m^{th} area total, is given by:

$$\widehat{Y}_m = y_m + \widehat{Y}_{mr} \tag{3.1}$$

where $y_m = \sum_{k=1}^{n_m} y_{mk}$: the sum of the sample values for the variable of interest in the m^{th} area; $\widehat{Y}_{mr} = \sum_{k=n_m+1}^{N_m} \widehat{y}_{mk}$ is a predictor of the total of the non-sampled units of the m^{th} area. The predictor $\widehat{y}_{mk}$ can be derived from a linear mixed model (LMM) (see Battese et al, 1988 [1] for unit level model, Fay and Herriot, 1979 [4] for area level model).

$$\log(y_{mk}) = \beta_0 + \boldsymbol{\beta}_1 \log(x_{mk}) + u_m + \epsilon_{mk}, \tag{3.2}$$

where y_{mk} and x_{mk} are the value of the k^{th} unit within the m^{th} area from the survey year and census year, respectively; u_m is the random effect of the m^{th} area, ε_{mk} is the error term. The log-transformation is applied to x_{mk} and y_{mk} to make the predictor and response variables approximately conform to normality. Then y_{mk} ($m = 1, ..., M$; $k = n_m + 1, ..., N_m$) is predicted using the inverse transformation

$$\widehat{y}_{mk} = \exp\left[\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} \log(x_{mk}) + \widehat{u}_m\right] \tag{3.3}$$

EBLUP: Empirical Best Linear Unbiased Predictor

$$\widehat{Y}_m^{\text{EBLUP}} = y_m + \sum_{k=n_m+1}^{N_m} \widehat{y}_{mk}$$

$\widehat{y}_{mk}$ is estimated using (3.3 and 3.2) assuming

$$u_m \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^2\right), \tag{3.4}$$

$$\epsilon_{mk} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{3.5}$$

The fixed effect $(\beta_0, \boldsymbol{\beta}_1)^T$ and the random effect $\widehat{u}_m$ are estimated via restricted maximum likelihood using the SAS Mixed procedure.

HB-T: Hierarchical Bayes estimator assuming *t*-distributed errors

$$\widehat{Y}_m^{\mathrm{HB-T}} = y_m + \sum_{k=n_m+1}^{N_m} \widehat{y}_{mk}$$

$\widehat{y}_{mk}$ is estimated using (3.3 and 3.2) assuming

$$u_m|\,\tau^2 \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^2\right), \tag{3.6}$$

$$\varepsilon_{mk}|\,\sigma^2 \overset{\text{iid}}{\sim} t\left(0, \sigma^2, \nu\right). \tag{3.7}$$

where u_m and ϵ_{mk} are independent. The parameter $(\beta_0, \boldsymbol{\beta}_1)^T$ and the random effect $\widehat{u}_m$ are estimated using a Markov Chain Monte Carlo (MCMC) procedure (see Trinh and Tran 2017 [14]) with the following specifications:

$$\begin{aligned}
\text{Parms}: \quad & \beta_1 = 0,\ \beta_2 = 1,\ \tau^2 = 1,\ \sigma^2 = 1 \\
\text{Priors}: \quad & \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \right) \\
& \tau^2,\ \sigma^2 \sim \text{igamma}(0.01,\ 0.01) \\
\text{Random}: \quad & u_m \sim \mathcal{N}(0, \tau^2) \\
\text{Likelihood}: \quad & \log(y_{mk})|\,u_m \sim t(\beta_1 + \beta_2 \log(x_{mk}) + u_m,\ \sigma^2,\ \nu)
\end{aligned}$$

HB-NN: Hierarchical Bayes estimator with error terms following a mixture of normal distributions

$$\widehat{Y}_m^{\mathrm{HB-NN}} = y_m + \sum_{k=n_m+1}^{N_m} \widehat{y}_{mk}$$

$\widehat{y}_{mk}$ is estimated using (3.3 and 3.2) assuming

$$u_m|\,\tau^2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2), \tag{3.8}$$

$$\varepsilon_{mk}|\,\sigma_1^2, \sigma_2^2, s_p \overset{\text{iid}}{\sim} (1 - s_p)\,\mathcal{N}\left(0, \sigma_1^2\right) + s_p\mathcal{N}\left(0, \sigma_2^2\right),\ \sigma_1 < \sigma_2, \tag{3.9}$$

$$s_p|\,p \overset{\text{iid}}{\sim} \text{Bin}\,(1; p)\,. \tag{3.10}$$

where u_m and ϵ_{mk} are independent. The parameter $(\beta_0, \boldsymbol{\beta}_1)^T$ and random effect $\widehat{u}_m$ are estimated using a SAS MCMC procedure (see Trinh and Tran 2017 [14]) with the following specifications:

$$\begin{aligned}
\text{Parms}: \quad & \beta_1 = 0,\ \beta_2 = 1,\ s_p = 0,\ \tau^2 = 0.01,\ \sigma_1^2 = 0.01,\ \sigma_2^2 = 1 \\
& p = \frac{1}{1 + \exp(-s_p)} \\
\text{Priors}: \quad & \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix} \right) \\
& \tau^2,\ \sigma_1^2,\ \sigma_2^2 \sim \text{igamma}(0.01,\ 0.01) \\
& s_p \sim \text{Bin}(1, p) \\
\text{Random}: \quad & u_m \sim \mathcal{N}(0, \tau^2),\ \mu = \beta_1 + \beta_2 \log(x_{mk}) + u_m, \\
& z_1 = \frac{\log(y_{mk}) - \mu}{\sigma_1},\ z_2 = \frac{\log(y_{mk}) - \mu}{\sigma_2} \\
\text{Likelihood}: \quad & \log(y_{mk})|\,u_m \sim \frac{p}{\sigma_1} \exp(-\frac{z_1^2}{2}) + \frac{1-p}{\sigma_2} \exp(-\frac{z_2^2}{2})
\end{aligned}$$

HB-TT: Hierarchical Bayes estimator with error terms following a mixture of t-distributions

$$\widehat{Y}_m^{\mathrm{HB-TT}} = y_m + \sum_{k=n_m+1}^{N_m} \widehat{y}_{mk}$$

$\widehat{y}_{mk}$ is estimated using (3.3 and 3.2) assuming

$$u_m | \tau^2 \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \tau^2\right), \tag{3.11}$$

$$\varepsilon_{mk} | \sigma_1^2, \sigma_2^2, s_p \overset{\text{iid}}{\sim} (1 - s_p)\, t\left(0, \sigma_1^2, \nu\right) + s_p t\left(0, \sigma_2^2, \nu\right), \ \sigma_1 < \sigma_2, \tag{3.12}$$

$$s_p | p \overset{\text{iid}}{\sim} \mathrm{Bin}\,(1; p)\,. \tag{3.13}$$

where u_m and ϵ_{mk} are independent. The parameter $(\beta_0, \boldsymbol{\beta}_1)^T$ and random effect $\widehat{u}_m$ are estimated using a SAS MCMC procedure with the following specifications:

$$\begin{aligned}
\text{Parms}: \quad & \beta_1 = 0,\ \beta_2 = 1,\ s_p = 0,\ \tau^2 = 0.01,\ \sigma_1^2 = 0.01,\ \sigma_2^2 = 1 \\
& p = \frac{1}{1 + \exp\left(-s_p\right)} \\
\text{Priors}: \quad & \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \mathrm{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix}\right) \\
& \tau^2,\ \sigma_1^2,\ \sigma_2^2 \sim \text{inverse gamma}(0.01,\ 0.01) \\
& s_p \sim \text{Binary}\,(1, p) \\
\text{Random}: \quad & u_m \sim \mathcal{N}\left(0, \tau^2\right),\ \mu = \beta_1 + \beta_2 \log\left(x_{mk}\right) + u_m, \\
& t_1 = \frac{\log\left(y_{mk}\right) - \mu}{\sigma_1},\ t_2 = \frac{\log\left(y_{mk}\right) - \mu}{\sigma_2} \\
\text{Likelihood}: \quad & \log\left(y_{mk}\right) | u_m \sim \frac{p}{\sigma_1}\left(1 + \frac{t_1^2}{\nu}\right)^{-\frac{1+\nu}{2}} + \frac{1-p}{\sigma_2}\left(1 + \frac{t_2^2}{\nu}\right)^{-\frac{1+\nu}{2}}
\end{aligned}$$

The degree of freedom is fixed ($\nu = 8$). The MCMC procedure discards the first $2,500$ as *burn-in* and keeps the next $12,500$ samples. The *thinning* rate of 5 is applied to produce $2,500$ *thinned* samples from the posterior distribution. Then u_m is predicted by the average of MCMC posterior estimates of u_m.

4 Application to ASPEP and ALFIN data

This research uses the ASPEP and ALFIN in California where there are 29 government functions in the ASPEP and 163 in the ALFIN..The performances of the five estimators (HT, EBLUP, HB-T, HB-NN, HB-TT) are evaluated within and across government functions using a “pairwise comparison” and a “percentile comparison”. The first comparison (pairwise comparison) compares the Relative Root Mean Square Errors (RRMSE) produced by the two competing estimators within each government function where an estimator is considered “better” than the other if it produces an estimate with smaller RRMSE in that area. In the second comparison (percentile comparison), a percentage level is set and then estimators are evaluated and compared based on the proportion of government functions producing estimates with RRMSEs smaller than the specified level.

4.1 Evaluation Design

The production sampling design is applied to select 1000 replicated samples from the 2012 CoG. Five estimators (HT, EBLUP, HB-T, HB-NN, HB-TT) are applied to produce estimates for the total Y_m. The quality of the estimators is evaluated using $\mathrm{RRMSE_m} = \sqrt{\frac{1}{\text{rep}} \sum_{i=1}^{\text{rep}} \left(\frac{\widehat{Y}_{m,i} - Y_m}{Y_m} \right)^2}$ where $\widehat{Y}_{m,i}$ is an estimate of Y_m, and $\text{rep} = 1,000$ is the number of replicate samples selected from the 2012 CoG data (CoG-E and CoG-F) and used to estimate totals for 2012 CoG data.

4.2 Application to ASPEP data

The parameter of interest in this study is the total number of full-time employees, Y_m, for each function code $m = 1, 2, ..., 29$ in the 2012 CoG-E of California. The 2007 CoG-E provides the auxiliary data used to estimate the 2012 CoG-E full-time employment using the HT, EBLUP, HB-T, HB-NN, and HB-TT estimators.

4.2.1 Pairwise Comparison

A pairwise comparison of the RRMSE of the HB-NN and HB-TT versus the HT, EBLUP, and HB-T estimators is given in Table 1. The values indicate the proportion that an estimator outperforms the other for RRMSE in 29 function codes in the ASPEP.

Table 1: Proportion that an Estimator Outperforms the Other for 29 Government Functions Tested in California for 2012 CoG-E

HT	HB-NN
$\frac{1}{29}$	$\frac{28}{29}$

EBLUP	HB-NN
$\frac{10}{29}$	$\frac{19}{29}$

HB-T	HB-NN
$\frac{11}{29}$	$\frac{17}{29}$

HT	HB-TT
$\frac{1}{29}$	$\frac{28}{29}$

EBLUP	HB-TT
$\frac{10}{29}$	$\frac{19}{29}$

HB-T	HB-TT
$\frac{10}{29}$	$\frac{18}{29}$

HB-NN	HB-TT
$\frac{13}{29}$	$\frac{15}{29}$

Data source: U.S. Census Bureau 2007 and 2012 CoG-E California

Among the seven pairwise comparisons in Table 1, the first three comparisons (Row 1 of Table 1) indicate that the HB-NN estimator performs better because it provides estimates with smaller RRMSEs. For the same reason, the last four comparisons (Rows 2 and 3 of Table 1) show the HB-TT estimator performs better when compared to the other estimators (HT, EBLUP, HB-T, HB-NN). The RRMSEs for each estimator are provided in Figure 1.

4.2.2 Percentile Comparison

The proportion of function codes (out of 29 in the ASPEP) where estimators produce estimates with RRMSEs less than 3 percent is given in Table 2.

Table 2: Proportion that Each Estimator Produces Relative Root Mean Square Error less than 3 Percent

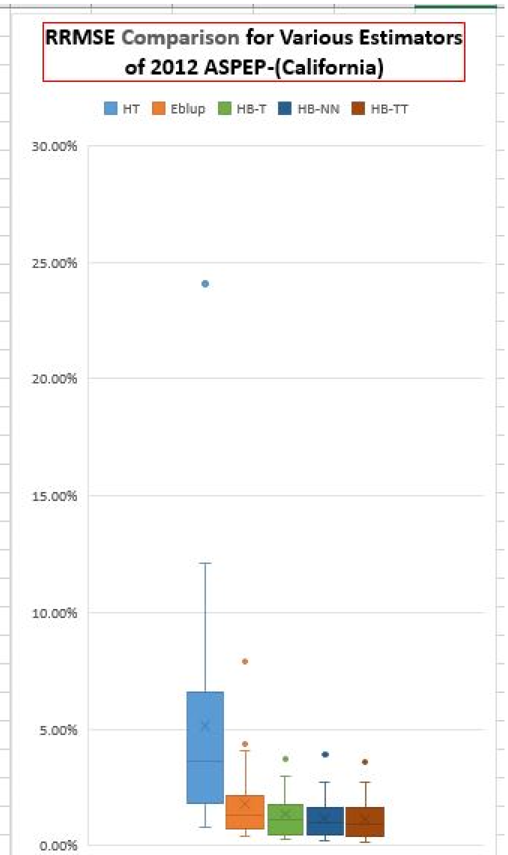
HT	EBLUP	HB-T	HB-NN	HB-TT
$\frac{13}{29}$	$\frac{24}{29}$	$\frac{26}{29}$	$\frac{28}{29}$	$\frac{28}{29}$

Data source: U.S. Census Bureau 2007 and 2012 CoG-E California

The HB-NN and HB-TT estimators produce estimates with RRMSEs less than 3 percent in more estimation areas (out of 29s in the ASPEP) than the HT, EBLUP and HB-T do. On average, ASPEP estimation accuracy (in terms of RRMSE) increases in the order of HT, EBLUP, HB-T, HB-NN and HB-TT (See Figure 1).

Figure 1: Relative Root Mean Square Error of Estimators and Their Box Plots

FC	HT	Eblup	HB-T	HB-NN	HB-TT
001	2.94%	0.52%	0.30%	0.29%	0.30%
005	0.95%	0.67%	0.71%	0.57%	0.56%
012	1.20%	1.30%	1.70%	1.65%	1.41%
016	11.09%	4.09%	3.79%	2.53%	2.44%
018	10.54%	4.35%	3.00%	2.50%	2.45%
023	1.95%	0.84%	0.77%	0.97%	0.86%
024	3.61%	0.83%	1.08%	1.20%	1.16%
025	0.88%	0.40%	0.49%	0.58%	0.62%
029	3.23%	1.20%	1.71%	1.49%	1.56%
032	1.16%	0.83%	0.34%	0.35%	0.37%
040	1.33%	0.63%	0.87%	0.78%	0.79%
044	3.82%	1.69%	0.42%	0.42%	0.33%
050	4.51%	3.50%	1.12%	0.69%	0.29%
052	4.04%	0.80%	1.10%	0.93%	0.90%
059	11.06%	7.91%	3.70%	3.92%	3.60%
061	10.75%	3.00%	2.81%	1.55%	1.65%
062	2.10%	0.66%	0.32%	0.39%	0.39%
079	0.78%	0.68%	0.32%	0.25%	0.24%
080	6.24%	1.81%	1.60%	1.25%	1.36%
081	6.69%	2.48%	1.41%	1.35%	1.43%
087	12.14%	1.35%	1.16%	1.18%	1.21%
089	5.22%	2.09%	2.73%	1.98%	2.08%
091	3.81%	2.19%	0.89%	0.38%	0.37%
092	2.46%	0.46%	0.50%	0.51%	0.49%
093	24.06%	1.54%	1.76%	1.80%	1.81%
094	2.39%	1.19%	1.11%	1.07%	1.10%
112	1.71%	1.50%	0.97%	0.23%	0.18%
124	6.44%	1.49%	2.66%	2.74%	2.70%
162	2.17%	1.26%	0.43%	0.50%	0.51%



Data source: U.S. Census Bureau 2007 and 2012 CoG-E California

4.3 Application to ALFIN data

The parameter of interest in this study is the total expenditures or revenues, Y_m, for each function code $m = 1, 2, ..., 163$ in the 2012 CoG-F of California. The 2007 CoG-F provides the auxiliary data used to estimate the 2012 CoG-F total expenditures or revenues using the HT, EBLUP, HB-T, HB-NN, and HB-TT estimators.

4.3.1 Pairwise Comparison

A pairwise comparison of the RRMSE of the HB-NN and HB-TT versus the HT, EBLUP, and HB-T estimators is given in Table 3. The values indicate the proportion that an estimator outperforms the other for RRMSE in 163 function codes in the ALFIN.

Table 3: Proportion that an Estimator Outperforms the Other for 163 Government Functions Tested in California for 2012 CoG-F

HT	HB-NN
$\frac{15}{163}$	$\frac{127}{163}$

EBLUP	HB-NN
$\frac{43}{163}$	$\frac{99}{163}$

HB-T	HB-NN
$\frac{63}{163}$	$\frac{79}{163}$

HT	HB-TT
$\frac{14}{163}$	$\frac{128}{163}$

EBLUP	HB-TT
$\frac{39}{163}$	$\frac{103}{163}$

HB-T	HB-TT
$\frac{63}{163}$	$\frac{79}{163}$

HB-NN	HB-TT
$\frac{59}{163}$	$\frac{83}{163}$

Data source: U.S. Census Bureau 2007 and 2012 CoG-F California

Among the seven pairwise comparisons in Table 3, the first three comparisons (Row 1 of Table 3) indicate that the HB-NN estimator performs better because it provides estimates with smaller RRMSEs. Similarly, the last four comparisons (Rows 2 and 3 of Table 3) show the HB-TT estimator performs better when compared to the other estimators (HT, EBLUP, HB-T, HB-NN).

4.3.2 Percentile Comparison

The proportion of function codes (out of 163 in the ALFIN) where estimators produce estimates with RRMSEs less than 3 percent and 5 percent are given in Table 4 and Table 5 respectively.

Table 4: Proportion that Each Estimator Produces Relative Root Mean Square Error less than 3 Percent

HT	EBLUP	HB-T	HB-NN	HB-TT
$\frac{45}{163}$	$\frac{75}{163}$	$\frac{90}{163}$	$\frac{90}{163}$	$\frac{93}{163}$

Data source: U.S. Census Bureau 2007 and 2012 CoG-F California

Table 5: Proportion that Each Estimator Produces Relative Root Mean Square Error less than 5 Percent

HT	EBLUP	HB-T	HB-NN	HB-TT
$\frac{68}{163}$	$\frac{95}{163}$	$\frac{104}{163}$	$\frac{105}{163}$	$\frac{105}{163}$

Data source: U.S. Census Bureau 2007 and 2012 CoG-F California

The HB-NN and HB-TT estimators produce estimates with RRMSEs less than 3 percent and 5 percent in more estimation areas (out of 163s in the ALFIN) than the HT, EBLUP and HB-T do. On average, ALFIN estimation accuracy (in terms of RRMSE) increases in the order of HT, EBLUP, HB-T, HB-NN and HB-TT.

5 Conclusion

5.1 HB-TT and HB-NN versus HT, EBLUP, HB-T estimators

Our comparisons show clearly superior performance by the two new Hierarchical Bayes (HB) estimators tested in this research for the surveys studied using the

ASPEP and ALFIN data for California (see Tables 1 and 3). To accommodate potential outliers in the ASPEP, Trinh and Tran (2017) [14] produced a robust HB estimator with error terms following a mixture of normal distributions, referred to as the HB-NN estimation method. In this research we also included a new HB estimator with error terms following a mixture of $t-$distributions which we call the HB-TT estimation method. Both the HB-NN and HB-TT demonstrated superior performance over the Horvitz-Thompson (HT) method, the Empirical Best Linear Unbiased Prediction (EBLUP) method, and the Hierarchical Bayes estimator with error terms following a $t-$distribution (HB-T) method (currently used at the U.S. Census Bureau), in terms of RRMSE when using the ASPEP and the ALFIN data. Using HB-NN and HB-TT, ASPEP estimation improvements are possible at the 3 percent level of RRMSE (see Table 2) and ALFIN estimation improvements are possible at the 3 percent and 5 percent levels of RRMSE (see Tables 4 and 5). The HB-TT estimator is slightly better than the HB-NN in terms of RRMSE (see Tables 1-5).

5.2 Limitations and Future Research

Further research is necessary to determine if similar improvements as those found in the particular government functions in California for these two surveys apply to other states. Similarly, we may want to consider how these estimators perform for other public sector and similar surveys. We plan to look for a more robust $t-$mixture using an optimal degree of freedom based on the akaike information criterion (AIC) or bayesian information criterion (BIC).

Acknowledgements

The authors thank Magdalena Ramos, Raemeka Mayo, Stephen Kaputa, and Carol Caldwell for careful reading of drafts of this paper and helpful comments.

Examples of Government Functions for the Surveys in this Research

FC	Description of Government Function for the ASPEP
001	Air Transportation
012	Elementary and Secondary - Instruction
018	Higher Education - Instructional
...	...

A complete description of Government Functions for the ASPEP can be found at http://census.gov/programs-surveys/apes.html

FC	Description of Government Function for the ALFIN
A01	Air Transportation Charges
A09	Elementary Secondary Education - Lunch Charges
A10	Elementary Secondary Education Tuition and Transportation Charges
...	...

A complete description of Government Functions for the ALFIN can be found at http://census.gov/programs-surveys/gov-finances.html

References

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

[2] Bassel, N. and Tran, B. (2017). A Hierarchical Bayesian Approach to Estimation for the Annual Survey of Local Government Finances. *American Statistical Association, JSM2017 Conference Proceedings.*

[3] Dumbacher, B. and Hill, C. (2014). Description of the 2014 Sample Design for the Annual Survey of Public Employment and Payroll. *Internal U.S. Census Bureau Memorandum. February 14, 2014*

[4] Fay, R. E. and Herriot, R. A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data, *Journal of American Statistical Association, 74, 269- 277*

[5] Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *International Society for Bayesian Analysis. 2006, No 3, pp. 515-533.*

[6] Lange, K.L., Little, R.J.A., and Taylor, J.M.G. (1989). Robust Statistical Modeling Using the t-distribution. *Journal of the American Statistical Association, 84(408), 881-896.*

[7] McLachlan, G. and Basford, K. (1988). Mixture Models: Inference and Applications to Clustering. *Vol. 84. Marcel Deker.*

[8] McLachlan, G. and Peel, D. (2000). Finite Mixture Models, *John Wiley and Sons.*

[9] McLachlan, G., Ng, S., and Bean, R. (2006). Robust Cluster Analysis via Mixture Models. *Austrian Journal of Statistics. Vol. 35 (2006), No 2-3, 157-174.*

[10] Ojo, O., Lougue, S., and Woldegerima, W. (2017). Bayesian Generalized Linear Mixed Modeling of Tuberculosis Using Informative Priors. *PLoS ONE, 12(3), e0172580. http://doi.org/10.1371/journal.pone.0172580*

[11] Peel, D. and McLachlan, G. (2000). Robust Mixture Modelling Using the $t-$Distribution, *Statistics and Computing (2000) 10, 339-348.*

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

[13] Trinh, G. and Tran, B. (2016). Outliers in Annual Survey of Public Employment & Payroll Small-area Approach, *American Statistical Association, JSM2016 Conference Proceedings.*

[14] Trinh, G. and Tran, B. (2017). Robust Estimation for the Annual Survey of Public Employment & Payroll Using Mixture of Linear Mixed-Effect Models with the MCMC Procedure, *American Statistical Association, JSM2017 Conference Proceedings.*