Hierarchical Bayesian Methods for Combining Surveys

Yang Cheng* Adrijo Chakraborty† Gauri Datta‡§

Abstract

In order to estimate the number of occupied households, US Census Bureau conducts many surveys. As a result, we get different estimates of the number of occupied households from these surveys. While each survey is useful, differences among the estimates they produce are sometimes very large. To resolve these differences, we propose in this study a hierarchical Bayesian method to obtain a more reliable estimate of the number occupied households by combining estimates from these surveys. Exploiting the repetitive nature of the surveys, we propose a time series model. We apply our method to the estimates from Current Population Survey (CPS)/Annual Social and Economic Supplement, CPS/Housing Vacancy Survey, American Community Survey, and American Housing Survey between 2002 and 2011 to produce a more reliable estimate of the number of occupied households. We implement our objective Bayesian method by Gibbs sampling.

Key Words: Current Population Survey, Gibbs Sampling, Noninformative Priors, Time Series

1. Introduction

One topic in the 2012 Federal Committee on Statistical Methodology (FCSM) Statistical Policy Seminar is about Dueling Official Statistics: minimizing differences through cross agency, understanding sources of differences, and reducing user confusion with joint data releases. In the Federal Government, many surveys produce different estimates for the same variable. For examples: Current Population Survey (CPS)/Annual Social and Economic Supplement (ASEC), CPS/Housing Vacancy Survey (HVS), American Community Survey (ACS), and American Housing Survey (AHS) produce the number of occupied households. Another example is that CPS/ASEC, CPS/HVS, and AHS produce vacancy rates. Every time, we release our survey reports. The public auditors may be confused when they see estimates are substantially different for the same variable across surveys. What is a reliable official statistics people can use? Hogan (2012) gave some suggestions on "Reducing User Confusion with Joint Data Releases and User Education". Cresce at el. (2013) proposed a residual calibration method to reduce the different estimates Household numbers among 4 surveys: CPS/ASEC, CPS/HVS, ACS, and AHS. In this paper, we will model one of most basic demographic concepts: households (occupied housing units) through hierarchical Bayesian method. In Table 1, we report the household estimates from 2002 to 2011,

*U.S. Census Bureau, Washington, DC 20233. This report is released to inform interested parties of research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

†University of Georgia, Athens, GA 30602

‡University of Georgia, Athens, GA 30602

§U.S. Census Bureau, Washington, DC 20233.

obtained by the Current Population Survey (CPS), the Housing Vacancy Survey (HVS), the American Community Survey (ACS) and the American Housing Survey (AHS). Differences among the survey estimates are noticeable in Table 1. Estimates obtained by the CPS are consistently high over the years and the estimates from the HVS and the AHS are typically low. In order to combine the estimates obtained by these surveys, we propose and discuss various hierarchical Bayesian (HB) models. In this paper, we study and compare the combined estimates obtained from these HB methods.

Table 1: *Estimates of households, obtained in different surveys (numbers in 1000s).*

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS/ASEC	111278	112000	113343	114384	116011	116783	117181	117538	119927	121084
HVS	104994	105636	106971	108667	109736	110173	110475	112295	112899	113533
ACS	107367	108420	109902	111091	111617	112378	113101	113616	114567	114992
AHS	.	105842	.	108871	.	110692	.	111806	.	114907

Table 2: *Standard errors of the estimates obtained in different surveys (numbers in 1000s).*

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS/ASEC	260	260	235	234	261	261	261	262	262	262
HVS	185	182	179	204	194	187	181	174	173	171
ACS	.	.	.	144	146	144	147	161	163	180
AHS	.	165	.	218	.	231	.	238	.	396

2. A Hierarchical Bayesian model to combine several unbiased survey estimates

Let θ be a population characteristic of interest and suppose estimates of θ are available from m different surveys. Moreover, suppose that these surveys are repeated over time, annually or biennially. Thus some surveys may not have been conducted over every time point of interest. While one or more surveys are conducted at every time point 1 to T, not all surveys are done at every time point. Suppose the i^{th} survey is conducted at time points belonging to a set $S_i \subset \{1, 2, \ldots, T\}$. We assume that, $\bigcup_{i=1}^{m} S_i = \{1, 2, \ldots, T\}$, i.e., at least one of the surveys is conducted every year. To estimate θ_t, the population characteristic of interest at time t, we consider the following model:

$$y_{it} = \theta_t + e_{it}, \qquad t \in S_i \subset \{1, 2, \ldots, T\}, \tag{2.1}$$

where, y_{it} is the estimate of θ_t from the i^{th} survey at the t^{th} time point. We assume that sampling errors $e_{it} \sim N(0, \sigma_{it}^2)$, $t \in S_i$, $i = 1\ldots, m$, are independently distributed. We assume that σ_{it}^2's are known. Now, we propose the following random walk model for θ_t:

$$\theta_t = \theta_{t-1} + e_t^*, \qquad t = 1, 2, \ldots, T, \tag{2.2}$$

where, e_t^*'s are independently distributed with a truncated normal distribution truncated above 0, with variance $\sigma_{e^*}^2$. We assume that $\sigma_{e^*}^2$ and θ_o are unknown. Our proposed hierarchical Bayesian model for estimating the number of households is:

$$\begin{aligned} \text{Model } M_1: \qquad y_{it}|\theta_0, \theta_t, \sigma_{e^*}^2 &\overset{\text{ind}}{\sim} N(\theta_t, \sigma_{it}^2),\ t \in S_i,\ i = 1, \ldots, m, \\ \theta_t &= \theta_{t-1} + e_t^*,\ t = 1, \ldots, T, \\ e_t^*|\sigma_{e^*}^2 &\overset{\text{iid}}{\sim} \text{truncated } N(0, \sigma_{e^*}^2), \end{aligned} \tag{2.3}$$

with lower truncation point 0. In model M_1, values of σ_{it}^2's are known. The values of θ_0 and $\sigma_{e^*}^2$ are not available, so we assign the following noninformative prior to those parameters: θ_o and $\sigma_{e^*}^2$ are independently distributed with Uniform$(0, \infty)$.

Since we assume improper prior to some parameters in the model, the propriety of the posterior distribution resulting from the model need to be ensured. Theorem 2.1 provides sufficient conditions for the propriety of the posterior density for the model stated above.

From Table 2 we see that standard errors are not available from the American Community Survey from $2002-2004$. Also, the American Housing Survey estimates along with the standard errors are missing at every alternative year from $2002-2011$ (Tables 1 and 2). Let us introduce indicator variables δ_{it}'s, such that, δ_{it} = 1 if data from the i^{th} survey is available at time t and δ_{it} = 0 otherwise, $i = 1\ldots, m$ and $t = 1, \ldots, T$. We also define, $n_t = \sum_{i=1}^{m} \delta_{it}$, $t = 1, \ldots, T$.

Theorem 2.1 *The posterior distribution resulting from model M_1 will be proper if (a) $n_t > 0$ for all t, and (b) the number of time points $T > 3$.*

Since there are some missing y_{it}'s along with σ_{it}^2's, we define the variable r such that, r_{it} = 0 if σ_{it}^2 is missing and $r_{it} = \frac{1}{\sigma_{it}^2}$ otherwise. The following full conditional distributions obtained below will be essential to perform a Gibbs Sampling.

(a) $\theta_T|\theta_0, \theta_1, \ldots, \theta_{T-1}, \sigma_{e^*}^2, y \sim$ truncated Normal with mean = $\dfrac{\sum_{i=1}^{m} r_{iT} y_{iT} + \sigma_{e^*}^{-2}\theta_{T-1}}{\sum_{i=1}^{m} r_{iT} + \sigma_{e^*}^{-2}}$

and variance = $(\sum_{i=1}^{m} r_{iT} + \sigma_{e^*}^{-2})^{-1}$ with lower truncation point θ_{T-1}.

(b) $\theta_t|\theta_0, \theta_1, ., \theta_{t-1}, \theta_{t+1}, ., \theta_T, \sigma_{e^*}^2, y \sim$ truncated Normal with

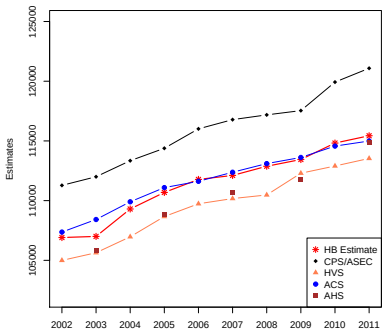


Figure 1: *Proposed HB estimates based on model M_1 and other survey estimates.*

mean = $\dfrac{\sum_{i=1}^{m} r_{it}y_{it} + \sigma_{e^*}^{-2}(\theta_{t-1}+\theta_{t+1})}{\sum_{i=1}^{m} r_{it} + 2\sigma_{e^*}^{-2}}$ and variance = $(\sum_{i=1}^{m} r_{it} + 2\sigma_{e^*}^{-2})^{-1}$ truncated in (θ_{t-1},θ_{t+1}); $t = 1, \ldots, T-1$.

(c) $\theta_0|\theta_1, \ldots, \theta_T, \sigma_{e^*}^2, y \sim$ truncated Normal with mean = θ_1 and variance = $\sigma_{e^*}^2$ truncated in (0,θ_1).

(d) $\sigma_{e^*}^2|\theta_0, \ldots, \theta_T, y \sim$ *Inverse-Gamma* (IG) with shape = $\frac{T}{2} - 1$, rate = $\sum_{t=1}^{T} \dfrac{(\theta_t - \theta_{t-1})^2}{2}$.

(If X $\sim$ *Inverse-Gamma*(α, β), then the pdf of X is $f(x) \propto x^{-\alpha-1}\exp\{-\frac{\beta}{x}\}$, where α is the shape and β is the rate parameter.)

We perform a Gibbs sampling to get the HB estimates of θ_t's. Table 4 presents the proposed HB estimates and posterior standard deviations of θ_t's. Figure 1 shows the proposed combined estimates and the survey estimates. From Table 3 we get the estimates of θ_0 and $\sigma_{e^*}^2$ obtained by our method. In Figure 3, we plot histograms based on the simulated values from the posterior distribution of $\sigma_{e^*}^2$. In Table 6 we provide some details about the simulated values from the posterior distribution of $\sigma_{e^*}^2$. Table 5 and Figure 2 (b) show that the posterior standard deviations associated with the Bayes estimates of θ_t obtained by our method are considerably lower than the standard errors of the survey estimates. This implies we achieve significant gain in precision by applying model M_1.

Table 3: *Summary of the posterior simulations (numbers in 1000s).*

Parameter	Posterior Mean	Posterior sd	Simulated Quantiles 2.5%	Median	97.5%
θ_0	105756.25	995.92	105296.47	105993.36	106489.93
$\sigma^2_{e^*}$	2369485.65	1921155.29	1300676.22	1854130.76	2774842.83

Table 4: *HB estimates and posterior standard deviations based on model M_1 (numbers in 1000s).*

Year	$\hat{\theta}_t$	Posterior sd	Year	$\hat{\theta}_t$	Posterior sd
2002	106909.21	103.48	2007	112110.28	92.93
2003	107002.75	93.73	2008	112877.75	103.76
2004	109300.26	141.15	2009	113443.00	97.42
2005	110688.17	94.65	2010	114823.40	107.55
2006	111775.22	103.65	2011	115433.41	107.08

2.1 Log Transformation

Since the values of household estimates are large and positive, we consider the following transformation: let, $y^*_{it} = log(y_{it})$ and $\theta^*_t = log(\theta_t)$. Now, we can rewrite equation (2.1) as,

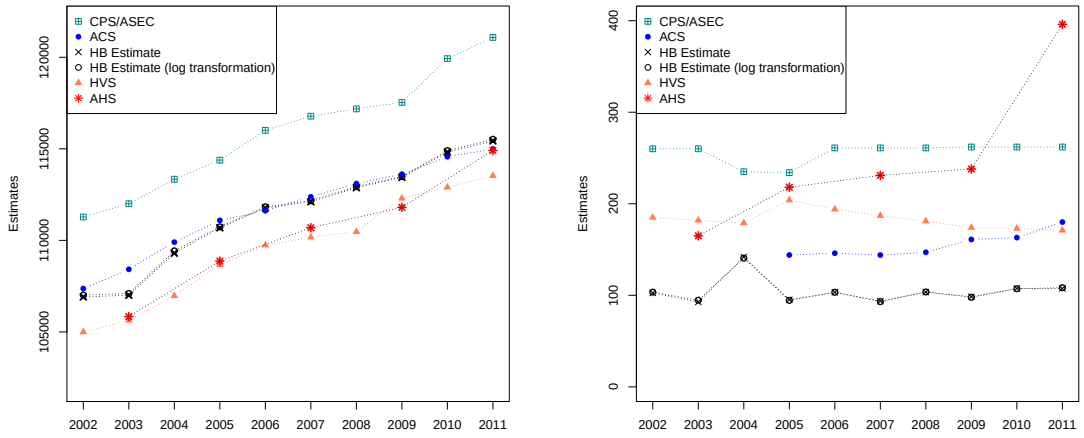
$$y^*_{it} = \theta^*_t + \epsilon_{it}, \qquad t \in S_i \subset \{1, 2, \ldots, T\}.$$

We assume that sampling errors $\epsilon_{it} \sim N(0, \tau^2_{it})$. Previously, we assumed that Var$(y_{it|\theta_t})$ = σ^2_{it}, where σ^2_{it}'s are known. Now, τ^2_{it} = Var$(y^*_{it}|\theta^*_t)$ = Var$(log(y_{it}|\theta_t)) \approx \dfrac{\sigma^2_{it}}{y^2_{it}}$, using Taylor series expansion. We obtain, the values of τ^2_{it}'s using this approximation. Similarly, as in equation (2.2), we assume,

$$\theta^*_t = \theta^*_{t-1} + \varepsilon_t, \qquad t = 1, 2, \ldots, T,$$

where, ε_t's are independently distributed with a truncated normal distribution truncated above 0, with variance σ^2_ε. We assume that σ^2_ε and θ^*_o are unknown. Now, with this re-parametrization, the proposed hierarchical Bayesian model in Section 2 could be rewritten as,

$$\begin{aligned} \text{Model } M_2: \qquad y^*_{it}|\theta^*_0, \theta^*_t, \sigma^2_\varepsilon &\overset{\text{ind}}{\sim} N(\theta^*_t, \tau^2_{it}),\ t \in S_i,\ i = 1, \ldots, m, \\ \theta^*_t &= \theta^*_{t-1} + \varepsilon_t,\ t = 1, \ldots, T, \\ \varepsilon_t|\sigma^2_\varepsilon &\overset{\text{iid}}{\sim} \text{truncated } N(0, \sigma^2_\varepsilon), \end{aligned} \tag{2.4}$$



(a) (b)

Figure 2: *(a) Proposed HB estimates based on models M_1 and M_2 (with and without log transformation) and other survey estimates. (b) Posterior standard deviations of the proposed HB estimates based on models M_1 and M_2, and the standard errors corresponding to the other survey estimates.*

with lower truncation point 0. We assume that, σ_ε^2 and θ_o^* are independently distributed with $\sigma_\varepsilon^2 \sim \text{Uniform}(0, \infty)$ and $\theta_o^* \sim \text{Uniform}(-\infty, \infty)$.

The resulting posterior distribution from this model will be proper if the sufficient conditions stated in Theorem 2.1 are satisfied. In order to estimate the parameters in this model, we use Gibbs sampling technique. Full conditional posterior distributions could be obtained by simple modifications of the full conditional distributions mentioned in Section 2. We run 5 chains and $10,000$ iterations for each chain. We discard first 50% observations of each chain and compute our estimates based on the remaining observations. Table 7 shows the summary of the posterior inference for σ_ε^2 . Histograms based on the posterior simulations for σ_ε^2 are shown in Figure 4.

Using the estimates of θ_t^* (say, $\hat{\theta}_t^*$), we can get the estimates of θ_t (say, $\hat{\theta}_t$) by the transformation $\hat{\theta}_t = \text{E}\left[\exp(\hat{\theta}_t^*)|y\right]$. In Table 8 we present the estimates of θ_t and the posterior standard deviations corresponding to the estimates. From Figure 2(a) we see that the estimates obtained by considering a log transformation almost coincide with the estimates obtained without considering a transformation. This applies to the posterior standard deviations as well (Figure 2(b)).

Table 5: *Posterior standard deviations and the standard errors (numbers in 1000s).*

Year	Proposed method (M_1) Posterior sd	CPS/ASEC s.e	HVS s.e	ACS s.e	AHS s.e
2002	103.48	260	185	.	.
2003	93.73	260	182	.	165
2004	141.15	235	179	.	.
2005	94.65	234	204	144	218
2006	103.65	261	194	146	.
2007	92.93	261	187	144	231
2008	103.76	261	181	147	.
2009	97.42	262	174	161	238
2010	107.55	262	173	163	.
2011	107.08	262	171	180	396

Table 6: *Details about the posterior simulations of $\sigma^2_{e^*}$.*

Simulated values of $\sigma^2_{e^*}\|y$	Proportion
$< 10^6$	0.10696
$10^6 - 5 \times 10^6$	0.82868
$5 \times 10^6 - 9 \times 10^6$	0.05176
$> 9 \times 10^6$	0.0126

Table 7: Summary of the posterior simulation for σ^2_{ε}

	Posterior	Posterior	Simulated Quantiles		
Parameter	Mean	sd	2.5%	Median	97.5%
σ^2_{ε}	0.00019	0.00015	0.00006	0.00015	0.00058

Table 8: *HB estimates and posterior standard deviations (numbers in 1000s).*

Year	$\hat{\theta}_t$	Posterior sd	Year	$\hat{\theta}_t$	Posterior sd
2002	107012.90	103.46	2007	112166.94	93.08
2003	107100.13	94.67	2008	112943.28	103.66
2004	109426.22	140.62	2009	113486.02	97.81
2005	110739.88	94.46	2010	114901.45	107.21
2006	111831.94	103.30	2011	115523.07	108.31

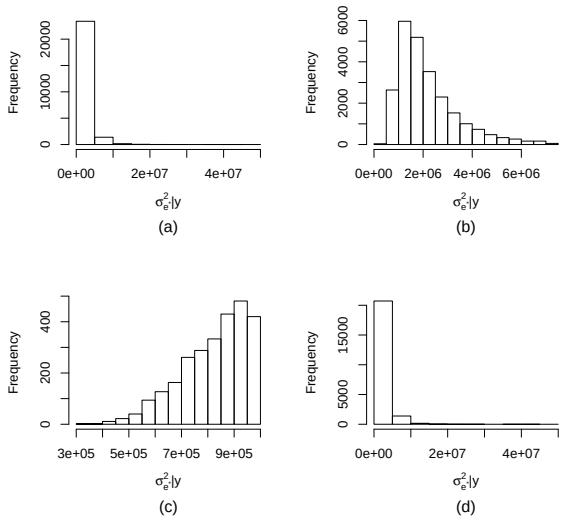
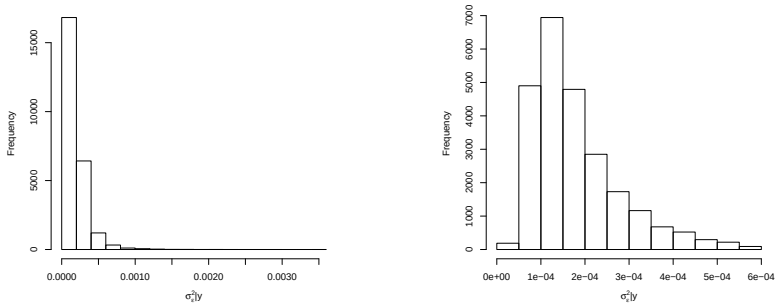


Figure 3: *Histograms of the posterior simulations for* $\sigma^2_{e^*}$ *(a) based an all simulated values (b) after dropping upper* 2.5% *observations (c) after dropping observations larger than* 10^6 *(d) after dropping observations smaller than* 10^6.



(a) (b)

Figure 4: *Histograms of the posterior simulations for* σ^2_ε *(a) based an all simulated values (b) after dropping upper* 2.5% *observations.*

3. Some new models accounting for sampling bias

From Figure 1 we see that there is a considerable difference in the survey estimates of households every year. To verify whether the surveys estimate the same quantity, we conduct hypothesis test to test the equality of $\mu_{it} = E(y_{it})$ among the surveys for each t, where y_{it} is the estimate of number of occupied household obtained from the i^{th} survey at the t^{th} year, for $i = 1 \ldots, m$ and $t = 1, \ldots, T$. For each year the null hypothesis that the surveys are estimating the same quantity was rejected convincingly. Motivated by this result we introduce a bias term for each survey modify model M_1 as follows:

$$\text{Model } M_3 \qquad y_{it}|\theta_0, \theta_t, \sigma^2_{e^*}, \alpha_i \overset{\text{ind}}{\sim} N(\theta_t + \alpha_i, \sigma^2_{it}), \tag{3.1}$$

$$\theta_t = \theta_{t-1} + e^*_t,$$

$$e^*_t|\sigma^2_{e^*} \overset{\text{iid}}{\sim} \text{truncated } N(0, \sigma^2_{e^*}), \tag{3.2}$$

with lower truncation point 0. Here, $t \in S_i$, $i = 1, \ldots, m$ and we impose the constraint $\sum_{i=1}^{m} \alpha_i = 0$, and assume a flat prior Uniform$(-\infty, \infty)$ for α_i, $i = 1, \ldots, m$. We assume the same priors for θ_0 and $\sigma^2_{e^*}$ as in model M_1. We describe the results based on the model M_3 in Table 9. Also, we provide the estimates of the contrasts of the bias parameters along with the posterior standard deviations in Table 11.

Table 9: *Estimates and posterior standard deviations deviations obtained from two different models (numbers in 1000s).*

	Estimates		Posterior SD	
Year	M_1	M_2	M_1	M_2
2002	106909.21	107127.72	103.48	156.28
2003	107002.75	107768.22	93.73	113.87
2004	109300.26	109125.61	141.15	146.15
2005	110688.17	110893.35	94.65	96.35
2006	111775.22	111824.62	103.65	111.76
2007	112110.28	112505.36	92.93	96.71
2008	112877.75	113016.36	103.76	106.61
2009	113443.00	113921.37	97.42	97.85
2010	114823.40	115029.86	107.55	110.01
2011	115433.41	115780.69	107.08	108.23

Population size and the number of occupied households have a natural relationship. That motivates us to use total population size as an auxiliary variable in order to improve the survey estimates. We assume that number of households and the total population size are linearly related. Let, x_t denote the population size at time point t. Our proposed model

considering x_t as an auxiliary variable is as follows:

$$\begin{aligned} y_{it} &= \theta_{it} + e_{it}, \\ \theta_{it} &= h_t + \alpha_i + b_{it}, \\ h_t &= \beta_0 + \beta x_t + \eta_t,\ t \in S_i,\ i = 1, \ldots, m, \end{aligned} \tag{3.3}$$

where α_i is the bias associated with the i^{th} survey, h_t is the true number of households at the year t. We impose an additive constraint $\sum_{i=1}^{m} \alpha_i = 0$ among the biases in the model. In the model, $e_{it} \sim N(0, D_{it})$, $b_{it} \sim N(0, \sigma_b^2)$, $\eta_t \sim N(0, \sigma_\eta^2)$, independently. The sampling variances D_{it}'s are known but the model variances σ_b^2 and σ_η^2 are unknown. Let, $\mu_i = \beta_0 + \alpha_i$, $i = 1, \ldots, m$. Since, $\sum_{i=1}^{m} \alpha_i = 0$, $\frac{1}{m} \sum_{i=1}^{m} \mu_i = \beta_0$.

We rewrite model (3.3) in the following form.

$$\begin{aligned} \text{Model } M_4 :\ & y_{it} | \alpha_i, h_t, b_{it} \sim N(h_t + \alpha_i + b_{it}, D_{it}),\ t \in S_i,\ i = 1, \ldots, m, \\ & h_t = \frac{1}{m} \sum_{i=1}^{m} \mu_i + \beta x_t + \eta_t, \\ & \eta_t | \sigma_\eta^2 \sim N(0, \sigma_\eta^2), \\ & b_{it} | \sigma_b^2 \sim N(0, \sigma_b^2),\ t = 1, \ldots, T,\ i = 1, \ldots, m, \\ & \pi(\mu, \beta, \sigma_b^2, \sigma_\eta^2) \propto 1, \\ & \frac{1}{m} \sum_{i=1}^{m} \mu_i = \beta_0, \text{ where } \mu_i = \beta_0 + \alpha_i. \end{aligned} \tag{3.4}$$

where, $\mu = (\mu_1, \ldots, \mu_m)^T$, $\beta \in \mathbf{R}$ and σ_b^2, $\sigma_\eta^2 \in \mathbf{R}^+$.

Theorem 3.1 *The posterior distribution resulting from Model M_4 will be proper if $T > 4$ and $m(T-1) > 5$.*

The joint pdf of $y, \mu, \beta, \eta, b, \sigma_b^2, \sigma_\eta^2$ from model M_4 is given by,

$$\begin{aligned} \pi(y, \mu, \beta, \eta, b, \sigma_b^2, \sigma_\eta^2) = C \times \exp\left\{ -\frac{1}{2} (y - Xw - Z_1\eta - b)^T D^{-1} (y - Xw - Z_1\eta - b) \right\} \\ \times \frac{1}{(\sigma_\eta^2)^{\frac{T}{2}}} \times \exp\left\{ -\frac{1}{2} \frac{\eta^T \eta}{2\sigma_\eta^2} \right\} \times \frac{1}{(\sigma_b^2)^{\frac{n}{2}}} \times \exp\left\{ -\frac{1}{2} \times \frac{b^T b}{2\sigma_b^2} \right\}, \end{aligned} \tag{3.5}$$

where, $Z_1 = \bigoplus_{t=1}^{T} 1_{n_t}$, $n_t = \sum_{i=1}^{m} = \delta_{it}$, where $\delta_{it} = 1$ if data from i^{th} survey is available at time t and $\delta_{it} = 0$ otherwise; $n = \sum_{t=1}^{T} n_t$. Here, $w = (\mu_1, \ldots, \mu_m, \beta)^T$ and, the design matrix is denoted by X.

We denote the identity matrix of order $n_t \times n_t$ by I_{n_t}.
In (3.5), $y = (y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_2 2}, \ldots, y_{n_T T})^T$, $\eta = (\eta_1, \ldots, \eta_T)^T$
$D = diag(D_{11}, \ldots, D_{n_1 1}, D_{12}, \ldots, D_{n_2 2}, \ldots, D_{n_T T})$,
$b = (b_{11}, \ldots, b_{n_1 1}, b_{12}, \ldots, b_{n_2 2}, \ldots, b_{n_T T})^T$.
Let us define, $f = y - Z_1\eta - b$, $g = y - Xw - b$ and $h = y - Xw - Z_1\eta$. The full conditional distributions obtained from (3.5) are given below,

(I) $w|y, \eta, b, \sigma_\eta^2, \sigma_b^2 \sim N((X^T D^{-1} X)^{-1} X^T D^{-1} f, (X^T D^{-1} X)^{-1})$,

(II) $\eta|y, w, b, \sigma_\eta^2, \sigma_b^2 \sim N\left((\sigma_\eta^{-2} I_T + Z_1^T D^{-1} Z_1)^{-1} Z_1^T D^{-1} g, (\sigma_\eta^{-2} I_T + Z_1^T D^{-1} Z_1)^{-1}\right)$,

(III) $b|y, w, \eta, \sigma_\eta^2, \sigma_b^2 \sim N\left((\sigma_b^{-2} I_n + D^{-1})^{-1} D^{-1} h, (\sigma_b^{-2} I_n + D^{-1})^{-1}\right)$,
where $n = (\sum_{t=1}^{T} n_t)$,

(IV) $\frac{1}{\sigma_\eta^2}|y, w, b, \eta, \sigma_b^2 \sim Gamma\left(\frac{T}{2} - 1, \frac{\eta^T \eta}{2}\right)$,

(V) $\frac{1}{\sigma_b^2}|y, w, b, \eta, \sigma_\eta^2 \sim Gamma\left(\frac{n}{2} - 1, \frac{b^T b}{2}\right)$.

We implement a Gibbs sampler using these conditional distributions. Estimates of h_t obtained from model M_4 and the standard deviation associated with the estimates from 2002 to 2011 are given in the first and third column of Table 10. From the fourth column of Table 10, we see that the posterior standard deviations associated with the estimates are on average larger than the sampling standard errors. This may be caused by using too many parameters in the model.

We consider another model which is almost same as Model M_4 but involves less number of parameters.

Model M_5 :
$$\begin{aligned} y_{it}|h_t, \alpha_i &\sim N(h_t + \alpha_i, D_{it}), \\ h_t &= \frac{1}{m}\sum_{i=1}^{m} \mu_i + \beta x_t + \eta_t, \\ \eta_t|\sigma_\eta^2 &\sim N(0, \sigma_\eta^2),\ t \in S_i,\ i = 1, \ldots, m, \\ \pi(\mu, \beta, \sigma_\eta^2) &\propto 1, \\ \frac{1}{m}\sum_{i=1}^{m} \mu_i &= \beta_0, \text{ where } \mu_i = \beta_0 + \alpha_i. \end{aligned} \tag{3.6}$$

Notation used in model M_5 has the same meaning as that defined before. The required full conditional distributions for model M_5 can be obtained with a little modification to the full conditional distributions corresponding to model M_4. We implement model M_5 and compute the estimates of number of households and the posterior standard deviations associated with the estimates given in Table 10. From Table 10 we see that while the posterior

standard deviations are considerably small for model M_5, the point estimates obtained using model M_5 are similar to the estimates obtained from model M_4 to a large extent. In Table 12 we estimate the bias for the surveys, where α_1 represents bias for CPS/ASEC, α_2 represents bias for HVS, α_3 is the bias for ACS and α_4 is the bias for AHS. In Table 13 we show the bias adjusted survey estimates.

Table 10: *HB estimates based on model M_4 and M_5 (numbers in 1000s)*

	Estimate		Posterior SD	
Year	M_4	M_5	M_4	M_5
2002	107156.26	107136.87	251.31	151.33
2003	107840.94	107786.69	211.06	112.93
2004	109140.83	109129.91	241.40	142.17
2005	110703.93	110869.47	183.58	94.77
2006	111730.94	111796.26	206.31	109.90
2007	112477.30	112495.83	172.28	95.77
2008	113046.61	113024.20	197.38	107.80
2009	113923.89	113934.36	180.77	97.56
2010	115131.95	115032.04	204.64	110.53
2011	115974.39	115788.58	188.92	108.81

Table 11: Bayesian inference of the bias contrasts based on M_3.

	Posterior	Posterior	Simulated Quantiles		
Parameter	Mean	sd	2.5%	Median	97.5%
$\alpha_1 - \alpha_2$	6388.732	99.34	6190.98	6389.906	6579.72
$\alpha_1 - \alpha_3$	4421.02	105.97	4218.93	4420.899	4630.89
$\alpha_1 - \alpha_4$	6123.44	136.46	5861.28	6122.919	6389.846
$\alpha_2 - \alpha_3$	-1967.72	86.29	-2135.48	-1967.01	-1801.206
$\alpha_2 - \alpha_4$	-265.2917	123.41	-508.40	-264.15	-20.36
$\alpha_3 - \alpha_4$	1702.41	126.35	1459.79	1701.92	1944.453

4. Summary

In this paper, we first observe the number of households estimated by different surveys differ considerably, which may create ambiguity among the researchers and impact decisions of government organizations. Secondly, we study various estimation methods through

Table 12: Bayesian inference of the bias contrasts based on M_5.

Parameter	Posterior Mean	Posterior sd	Simulated Quantiles 2.5%	Median	97.5%
$\alpha_1 - \alpha_2$	6390.31	98.44	6200.86	6390.38	6583.22
$\alpha_1 - \alpha_3$	4416.34	102.90	4216.10	4416.27	4618.69
$\alpha_1 - \alpha_4$	6127.05	137.17	5859.32	6127.30	6395.12
$\alpha_2 - \alpha_3$	-1973.98	86.61	-2141.1	-1975.04	-1803.76
$\alpha_2 - \alpha_4$	-263.27	125.39	-505.33	-262.62	-12.56
$\alpha_3 - \alpha_4$	1710.71	128.53	1463.46	1710.32	1962.27

different models to combine estimates from different surveys. Model 1 assumes all the surveys are unbiased. Model 2 uses Log transformation. Model 3 is a weighted average model. Model 4 accounts for potential biases among the surveys and replaces the Random Walk by a linear regression. Model 5 is almost same as Model 4, but involves less number of parameters. Our proposed methods successfully combine the survey estimates, which could be helpful to the researchers. Finally, we considered bias in the model and performed an exploratory analysis.We have shown that considerable gain in terms of precision can be achieved using some of these methods, specially, models 3 and 5 fit our data better because of survey bias. We should evaluate the precision by mean square errors instead of variance.

Table 13: *Bias corrected estimates of households from* $2002 - 2011$ *for three different surveys based on model* M_5 *(numbers in 1000s).*

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS/ASEC	107044.58	107766.58	109109.58	110150.58	111777.58	112549.58	112947.58	113304.58	115693.58	116850.58
HVS	107150.89	107792.89	109127.89	110823.89	111892.89	112329.89	112631.89	114451.89	115055.89	115689.89
ACS	107549.91	108602.91	110084.91	111273.91	111799.91	112560.91	113283.91	113798.91	114749.91	115174.91
AHS	.	107735.62	.	110764.62	.	112585.62	.	113699.62	.	116800.62

REFERENCES

Arthur Cresce Jr, Yang Cheng, and Christopher Grieves (2013). Household Estimates Conundrum: Effort to Develop More Consistent Household Estimates Across Current Survey. *2013 Federal Committee on Statistical Methodology Research Conference.*

Alan Gelfand and Adrian Smith (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association,* 85, (410) 398-409.

Howard R. Hogan (2012). Reducing User Confusion with Joint Data Releases and User Education. *2012 Federal Committee on Statistical Methodology Statistical Policy Seminar.*