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).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CPS/ASEC</td>
<td>111278</td>
<td>112000</td>
<td>113343</td>
<td>114384</td>
<td>116011</td>
<td>116783</td>
<td>117181</td>
<td>117538</td>
<td>119927</td>
<td>121084</td>
</tr>
<tr>
<td>HVS</td>
<td>104994</td>
<td>105636</td>
<td>106971</td>
<td>108667</td>
<td>109736</td>
<td>110173</td>
<td>110475</td>
<td>112295</td>
<td>112899</td>
<td>113533</td>
</tr>
<tr>
<td>ACS</td>
<td>107367</td>
<td>108420</td>
<td>109902</td>
<td>111091</td>
<td>111617</td>
<td>112378</td>
<td>113101</td>
<td>113616</td>
<td>114567</td>
<td>114902</td>
</tr>
<tr>
<td>AHS</td>
<td>105842</td>
<td>108871</td>
<td>110692</td>
<td>111806</td>
<td>114907</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CPS/ASEC</td>
<td>260</td>
<td>260</td>
<td>235</td>
<td>234</td>
<td>261</td>
<td>261</td>
<td>261</td>
<td>262</td>
<td>262</td>
<td>262</td>
</tr>
<tr>
<td>HVS</td>
<td>185</td>
<td>182</td>
<td>179</td>
<td>204</td>
<td>194</td>
<td>187</td>
<td>181</td>
<td>174</td>
<td>173</td>
<td>171</td>
</tr>
<tr>
<td>ACS</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>144</td>
<td>146</td>
<td>144</td>
<td>147</td>
<td>161</td>
<td>163</td>
</tr>
<tr>
<td>AHS</td>
<td>.</td>
<td>165</td>
<td>.</td>
<td>218</td>
<td>.</td>
<td>231</td>
<td>.</td>
<td>238</td>
<td>.</td>
<td>396</td>
</tr>
</tbody>
</table>

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

Let \( \theta \) be a population characteristic of interest and suppose estimates of \( \theta \) 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 \( \theta_t \), the population characteristic of interest at time \( t \), we consider the following model:

\[
y_{it} = \theta_t + e_{it}, \quad t \in S_i \subset \{1, 2, \ldots, T\},
\]

(2.1)
where, \( y_{it} \) is the estimate of \( \theta_t \) 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 \( \sigma_{it}^2 \)'s are known. Now, we propose the following random walk model for \( \theta_t \):

\[
\theta_t = \theta_{t-1} + e_t^*, \quad t = 1, 2, \ldots, T, \quad (2.2)
\]

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 \( \theta_0 \) are unknown. Our proposed hierarchical Bayesian model for estimating the number of households is:

Model \( M_1 \):

\[
y_{it} | \theta_0, \theta_t, \sigma_{e^*}^2 \overset{\text{ind}}{\sim} N(\theta_t, \sigma_{it}^2), \quad t \in S_i, \quad i = 1, \ldots, m, \\
\theta_t = \theta_{t-1} + e_t^*, \quad t = 1, \ldots, T; \\
e_t^* | \sigma_{e^*}^2 \overset{\text{iid}}{\sim} \text{truncated} \ N(0, \sigma_{e^*}^2), \quad (2.3)
\]

with lower truncation point 0. In model \( M_1 \), values of \( \sigma_{it}^2 \)'s are known. The values of \( \theta_0 \) and \( \sigma_{e^*}^2 \) are not available, so we assign the following noninformative prior to those parameters: \( \theta_0 \) 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 \( \delta_{it} \)'s, such that, \( \delta_{it} = 1 \) if data from the \( i^{th} \) survey is available at time \( t \) and \( \delta_{it} = 0 \) 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 \( \sigma_{it}^2 \)'s, we define the variable \( r \) such that, \( r_{it} = 0 \) if \( \sigma_{it}^2 \) 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) \( \theta_T | \theta_0, \theta_1, \ldots, \theta_{T-1}, \sigma_{e^*}^2, y \sim \text{truncated Normal with mean} = \frac{\sum_{i=1}^{m} r_{iT} y_{iT} + \sigma_{e^*}^2 \theta_{T-1}}{\sum_{i=1}^{m} r_{iT} + \sigma_{e^*}^{-2}} \)

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

(b) \( \theta_t | \theta_0, \theta_1, \ldots, \theta_{t-1}, \theta_{t+1}, \ldots, \theta_T, \sigma_{e^*}^2, y \sim \text{truncated Normal with} \)
mean = \frac{\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}} \quad \text{and variance} = \frac{(\sum_{i=1}^{m} r_{it} + 2\sigma_{e}^{-2})^{-1}}{\text{truncated in } (\theta_{t-1}, \theta_{t+1}); \ t = 1, \ldots, T - 1.}

(c) \theta_{0}|\theta_{1}, \ldots, \theta_{T}, \sigma_{e}^{2}, y \sim \text{truncated Normal with mean } = \theta_{1} \text{ and variance } = \sigma_{e}^{2}, \text{ truncated in } (0, \theta_{1}).

(d) \sigma_{e}^{2}|\theta_{0}, \ldots, \theta_{T}, y \sim \text{Inverse-Gamma (IG) with shape } = \frac{T}{2} - 1, \text{ rate } = \frac{T}{2} \sum_{t=1}^{T} (\theta_{t} - \theta_{t-1})^{2} / 2.

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

We perform a Gibbs sampling to get the HB estimates of \theta_{t}'s. Table 4 presents the proposed HB estimates and posterior standard deviations of \theta_{t}'s. Figure 1 shows the proposed combined estimates and the survey estimates. From Table 3 we get the estimates of \theta_{0} \text{ and } \sigma_{e}^{2}, \text{ obtained by our method. In Figure 3, we plot histograms based on the simulated values from the posterior distribution of } \sigma_{e}^{2}. \text{ In Table 6 we provide some details about the simulated values from the posterior distribution of } \sigma_{e}^{2}. \text{ Table 5 and Figure 2 (b) show that the posterior standard deviations associated with the Bayes estimates of } \theta_{t}, \text{ 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}.\n
Figure 1: Proposed HB estimates based on model } M_{1} \text{ and other survey estimates.}
Table 3: Summary of the posterior simulations (numbers in 1000s).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>Posterior sd</th>
<th>Simulated Quantiles 2.5%</th>
<th>Median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_0$</td>
<td>105756.25</td>
<td>995.92</td>
<td>105296.47</td>
<td>105993.36</td>
<td>106489.93</td>
</tr>
<tr>
<td>$\sigma^2_\epsilon$</td>
<td>2369485.65</td>
<td>1921155.29</td>
<td>1300676.22</td>
<td>1854130.76</td>
<td>2774842.83</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Year</th>
<th>$\theta_t$ Posterior Mean</th>
<th>$\theta_t$ Posterior sd</th>
<th>Year</th>
<th>$\theta_t$ Posterior Mean</th>
<th>$\theta_t$ Posterior sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>106909.21</td>
<td>103.48</td>
<td>2007</td>
<td>112110.28</td>
<td>92.93</td>
</tr>
<tr>
<td>2003</td>
<td>107002.75</td>
<td>93.73</td>
<td>2008</td>
<td>112877.75</td>
<td>103.76</td>
</tr>
<tr>
<td>2004</td>
<td>109300.26</td>
<td>141.15</td>
<td>2009</td>
<td>113443.00</td>
<td>97.42</td>
</tr>
<tr>
<td>2005</td>
<td>110688.17</td>
<td>94.65</td>
<td>2010</td>
<td>114823.40</td>
<td>107.55</td>
</tr>
<tr>
<td>2006</td>
<td>111775.22</td>
<td>103.65</td>
<td>2011</td>
<td>115433.41</td>
<td>107.08</td>
</tr>
</tbody>
</table>

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,

$$y_{it}^* = \theta_t^* + \epsilon_{it}, \quad t \in S_i \subset \{1, 2, \ldots, T\}.$$  

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

$$\theta_t^* = \theta_{t-1}^* + \epsilon_t, \quad t = 1, 2, \ldots, T,$$

where $\epsilon_t$’s are independently distributed with a truncated normal distribution truncated above 0, with variance $\sigma^2_{\epsilon}$. We assume that $\sigma^2_{\epsilon}$ and $\theta_o^*$ are unknown. Now, with this re-parametrization, the proposed hierarchical Bayesian model in Section 2 could be rewritten as,

Model $M_2$ :  

$$y_{it}^*|\theta_0^*, \theta_t^*, \sigma^2_{\epsilon} \text{ iid } N(\theta_t^*, \tau^2_{it}), \quad t \in S_i, \quad i = 1, \ldots, m,$$

$$\theta_t^* = \theta_{t-1}^* + \epsilon_t, \quad t = 1, \ldots, T,$$

$$\epsilon_t|\sigma^2_{\epsilon} \text{ iid } \text{ truncated } N(0, \sigma^2_{\epsilon}),$$

(2.4)
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, $\sigma^2_\varepsilon$ and $\theta^*_t$ are independently distributed with $\sigma^2_\varepsilon \sim \text{Uniform}(0, \infty)$ and $\theta^*_t \sim \text{Uniform}(-\infty, \infty)$.

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 $\sigma^2_\varepsilon$. Histograms based on the posterior simulations for $\sigma^2_\varepsilon$ are shown in Figure 4.

Using the estimates of $\theta^*_t$ (say, $\hat{\theta}_t^*$), we can get the estimates of $\theta_t$ (say, $\hat{\theta}_t$) by the transformation $\hat{\theta}_t = E \left[ \exp(\hat{\theta}_t^*) | y \right]$. In Table 8 we present the estimates of $\theta_t$ 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).

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

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

| Simulated values of $\sigma^2_{\varepsilon}|y$ | 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 $\sigma^2_{\varepsilon}$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>Posterior sd</th>
<th>Simulated Quantiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_{\varepsilon}$</td>
<td>0.00019</td>
<td>0.00015</td>
<td>0.00006  0.00015  0.00058</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Year</th>
<th>$\theta_1$</th>
<th>Posterior sd</th>
<th>Year</th>
<th>$\theta_2$</th>
<th>Posterior sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>107012.90</td>
<td>103.46</td>
<td>2007</td>
<td>112166.94</td>
<td>93.08</td>
</tr>
<tr>
<td>2003</td>
<td>107100.13</td>
<td>94.67</td>
<td>2008</td>
<td>112943.28</td>
<td>103.66</td>
</tr>
<tr>
<td>2004</td>
<td>109426.22</td>
<td>140.62</td>
<td>2009</td>
<td>113486.02</td>
<td>97.81</td>
</tr>
<tr>
<td>2005</td>
<td>110739.88</td>
<td>94.46</td>
<td>2010</td>
<td>114901.45</td>
<td>107.21</td>
</tr>
<tr>
<td>2006</td>
<td>111831.94</td>
<td>103.30</td>
<td>2011</td>
<td>115523.07</td>
<td>108.31</td>
</tr>
</tbody>
</table>
Figure 3: Histograms of the posterior simulations for $\sigma_e^2$ (a) based on all simulated values (b) after dropping upper 2.5% observations (c) after dropping observations larger than $10^6$ (d) after dropping observations smaller than $10^6$.

Figure 4: Histograms of the posterior simulations for $\sigma_e^2$ (a) based on 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:

Model \( M_3 \)

\[
\begin{align*}
    y_{it} | \theta_0, \theta_t, \sigma^2_{e^*}, \alpha_i &\sim N(\theta_t + \alpha_i, \sigma^2_{it}), \\
    \theta_t &= \theta_{t-1} + e^*_t, \\
    e^*_t | \sigma^2_{e^*} &\sim \text{truncated } N(0, \sigma^2_{e^*}),
\end{align*}
\]

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 \( \alpha_i, i = 1, \ldots, m \). We assume the same priors for \( \theta_0 \) 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).

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

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:

\[
y_{it} = \theta_{it} + e_{it},
\]

\[
\theta_{it} = h_t + \alpha_i + b_{it},
\]

\[
h_t = \beta_0 + \beta x_t + \eta_t, \quad t \in S_i, \quad i = 1, \ldots, m, \tag{3.3}
\]

where \( \alpha_i \) 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 \( \sigma_b^2 \) and \( \sigma_\eta^2 \) are unknown. Let, \( \mu_i = \beta_0 + \alpha_i, i = 1, \ldots, m \). Since, \( \frac{1}{m} \sum_{i=1}^{m} \alpha_i = 0 \), \( \frac{1}{m} \sum_{i=1}^{m} \mu_i = \beta_0 \).

We rewrite model (3.3) in the following form.

**Model M4**: \( y_{it}|\alpha_i, h_t, b_{it} \sim N(h_t + \alpha_i + b_{it}, D_{it}), \quad t \in S_i, \quad i = 1, \ldots, m, \)

\[ h_t = \frac{1}{m} \sum_{i=1}^{m} \mu_i + \beta x_t + \eta_t, \]

\[ \eta_t|\sigma_\eta^2 \sim N(0, \sigma_\eta^2), \]

\[ b_{it}|\sigma_b^2 \sim N(0, \sigma_b^2), \quad t = 1, \ldots, T, \quad i = 1, \ldots, m, \]

\[ \pi(\mu, \beta, \sigma_b^2, \sigma_\eta^2) \propto 1, \]

\[ \frac{1}{m} \sum_{i=1}^{m} \mu_i = \beta_0, \quad \text{where} \quad \mu_i = \beta_0 + \alpha_i. \tag{3.4} \]

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

**Theorem 3.1** The posterior distribution resulting from Model M4 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 M4 is given by,

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

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), \(y = (y_1, \ldots, y_{n_1}, y_{12}, \ldots, y_{n_2}, \ldots, y_{n_T})^T\), \(\eta = (\eta_1, \ldots, \eta_T)^T\)
\(D = \text{diag}(D_{11}, \ldots, D_{n_1}, D_{12}, \ldots, D_{n_2}, \ldots, D_{n_T})\),
\(b = (b_{11}, \ldots, b_{n_1}, b_{12}, \ldots, b_{n_2}, \ldots, b_{n_T})^T\).

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) \(w | y, \eta, b, \sigma_y^2, \sigma_b^2 \sim N((X^T D^{-1} X)^{-1} X^T D^{-1} f, (X^T D^{-1} X)^{-1})\),
- (II) \(\eta | y, w, b, \sigma_y^2, \sigma_b^2 \sim N((\sigma_y^2 I_T + Z_1^T D^{-1} Z_1)^{-1} Z_1^T D^{-1} g, (\sigma_y^2 I_T + Z_1^T D^{-1} Z_1)^{-1})\),
- (III) \(b | y, w, \eta, \sigma_y^2, \sigma_b^2 \sim N((\sigma_b^2 I_n + D^{-1} D^{-1} h, (\sigma_b^2 I_n + D^{-1} D^{-1})^{-1})\),

where \(n = (\sum_{t=1}^{n_T} n_t)\).

- (IV) \(\frac{1}{\sigma^2_\eta} | y, w, b, \eta \sim Gamma \left(\frac{T}{2} - 1, \frac{\eta^T \eta}{2}\right)\),
- (V) \(\frac{1}{\sigma^2_b} | y, w, b, \eta \sim Gamma \left(\frac{n}{2} - 1, \frac{\eta^T \eta}{2}\right)\).

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 \(M_5\):**

\[ y_{it} | h_t, \alpha_i \sim N(h_t + \alpha_i, D_{it}) , \]

\[ h_t = \frac{1}{m} \sum_{i=1}^{m} \mu_i + \beta x_t + \eta_t, \]

\[ \eta_t | \sigma^2_\eta \sim N(0, \sigma^2_\eta), \quad t \in S_i, \quad i = 1, \ldots, m, \]

\[ \pi(\mu, \beta, \sigma^2_\eta) \propto 1, \]

\[ \frac{1}{m} \sum_{i=1}^{m} \mu_i = \beta_0, \quad \text{where} \quad \mu_i = \beta_0 + \alpha_i. \quad (3.6) \]

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 $\alpha_1$ represents bias for CPS/ASEC, $\alpha_2$ represents bias for HVS, $\alpha_3$ is the bias for ACS and $\alpha_4$ 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)

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

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

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

### 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 $M_5$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>Posterior sd</th>
<th>Simulated Quantiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1 - \alpha_2$</td>
<td>6390.31</td>
<td>98.44</td>
<td>6200.86</td>
</tr>
<tr>
<td>$\alpha_1 - \alpha_3$</td>
<td>4416.34</td>
<td>102.90</td>
<td>4216.10</td>
</tr>
<tr>
<td>$\alpha_1 - \alpha_4$</td>
<td>6127.05</td>
<td>137.17</td>
<td>5859.32</td>
</tr>
<tr>
<td>$\alpha_2 - \alpha_3$</td>
<td>-1973.98</td>
<td>86.61</td>
<td>-2141.1</td>
</tr>
<tr>
<td>$\alpha_2 - \alpha_4$</td>
<td>-263.27</td>
<td>125.39</td>
<td>-505.33</td>
</tr>
<tr>
<td>$\alpha_3 - \alpha_4$</td>
<td>1710.71</td>
<td>128.53</td>
<td>1463.46</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CPS/ASEC</td>
<td>107044.58</td>
<td>107766.58</td>
<td>109109.58</td>
<td>110150.58</td>
<td>111777.58</td>
<td>112549.58</td>
<td>112947.58</td>
<td>113304.58</td>
<td>115693.58</td>
<td>116850.58</td>
</tr>
<tr>
<td>HVS</td>
<td>107150.89</td>
<td>107792.89</td>
<td>109127.89</td>
<td>110823.89</td>
<td>111892.89</td>
<td>112329.89</td>
<td>112631.89</td>
<td>11451.89</td>
<td>115055.89</td>
<td>115689.89</td>
</tr>
<tr>
<td>ACS</td>
<td>107549.91</td>
<td>108602.91</td>
<td>110084.91</td>
<td>111773.91</td>
<td>111799.91</td>
<td>112560.91</td>
<td>113283.91</td>
<td>113798.91</td>
<td>114749.91</td>
<td>115174.91</td>
</tr>
</tbody>
</table>

REFERENCES

