A Hierarchical Bayesian Approach to Estimation for the Annual Survey of Public Employment & Payroll

Brian Dumbacher[1], Michael D. Larsen[2]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
[2]The George Washington University, Washington, DC 20052

Abstract
The Annual Survey of Public Employment & Payroll (ASPEP) is conducted by the U.S. Census Bureau to collect data on federal, state, and local government civilian employees. Estimates of local government totals are calculated for domains created by crossing state and government function, where functions range from air transportation to water supply. To calculate estimates at such a detailed level, the Census Bureau uses small area methods that borrow strength from other domains through auxiliary information from the most recent Census of Governments. In this paper, we study the properties of the composite estimator used during ASPEP's 2009 sample design and explore a new hierarchical Bayesian approach to small area estimation. We consider various models and investigate model diagnostics.

Key Words: Small area estimation; Composite estimator; Hierarchical Bayes; Model validation; Government units

1. Introduction

Every five years, in years ending in "2" and "7," the U.S. Census Bureau conducts the Census of Governments (CoG), which collects data on the organization, finances, and employment of the approximately 90,000 governments in the United States. The data collected are important in measuring the public-sector component of the economy, developing public policy, and understanding relationships among governments. During intercensal years, annual sample surveys are conducted to collect similar information.

One such survey is the Annual Survey of Public Employment & Payroll (ASPEP), which collects data on the number and pay of government civilian employees. ASPEP is made up of three components: a census of select federal agencies, a census of the 50 state governments, and a two-phase, stratified, probability-proportional-to-size sample of around 10,500 local governments (U.S. Census Bureau, 2014). About two years after each CoG, a new sample of local governments is selected. For example, the 2009 sample design is based on the 2007 CoG, and the current 2014 sample design is based on the 2012 CoG.

Estimates of local government totals are calculated for domains created by crossing state and government function. To calculate estimates at such a detailed level, the Census Bureau uses small area methods that borrow strength from similar domains through auxiliary information from the most recent CoG. The composite estimator used during ASPEP's 2009 sample design is based on an implicit model and equals a weighted average of direct and synthetic estimators (Tran and Cheng, 2011). In this paper, we study the design-based properties of the composite estimator and explore explicit models that try to address potential areas for improvement. The hierarchical Bayesian approach to modeling, which is new in application to ASPEP, offers a convenient setting in which to fit and evaluate models.

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

2. Composite Methodology

2.1 Domains

The domains of interest are created by crossing state f and function d. There are 49 states (Hawaii and the District of Columbia are excluded because censuses of local governments are conducted there), 29 functions, and a total of 1,421 (= 49 $\times$ 29) domains, although some domains may represent structural zeroes. A complete list of functions and corresponding codes can be found in Appendix A. The key study variables are the number of full-time employees (*ftemp*), full-time pay (*ftpay*), the number of part-time employees (*ptemp*), part-time pay (*ptpay*), and the number of part-time hours (*pthours*). For domain fd, the estimand is the total of study variable y, which is denoted Y_{fd}. To illustrate the structure of the data, Table 1 provides an example for East Wenatchee municipal government in Washington.

Table 1: Data Example for East Wenatchee Municipal Government

Function	*ftemp07*	*ftpay07*	*ptemp07*	*ptpay07*	*pthours07*
023	2	8,346	0	0	0
025	3	11,947	2	921	83
029	1	4,298	9	8,843	483
044	8	31,306	1	1,403	79
050	4	16,160	0	0	0
062	20	98,752	0	0	0
162	3	10,129	0	0	0

Note: For information on sampling and non sampling error or definitions, see
<www.census.gov/govs/apes/how_data_collected.html>
Source: U.S. Census Bureau, 2007 Census of Governments: Employment

2.2 Small Area Concepts

Small area estimation deals with calculating estimates and corresponding measures of variability for areas, or domains, whose sample sizes are too small for reliable direct estimation. Small area methods effectively increase the sample size by "borrowing strength" from similar domains through models and auxiliary data. Models can be implicit or explicit in how they do this, but in either case the goal is an appreciable increase in estimation accuracy over that of the direct estimator.

To achieve this goal, many estimators in the literature take the composite form of a weighted average of direct and synthetic estimators. Direct estimators, such as the Horvitz-Thompson estimator, typically have small bias and large variance, whereas synthetic estimators (Gonzalez, 1973) typically have large bias and small variance. The composite form tries to balance the variability of the direct estimator against the bias of the synthetic estimator. A comprehensive account of small area estimation is given in Rao (2003), and recent developments are discussed in Pfeffermann (2013).

Small area methods are used for ASPEP because the element sample sizes by domain cannot be controlled and may be too small for reliable direct estimation. This follows mainly from the fact that the sample design is not a direct-element design. Rather, the sampling units are governments, which are clusters of functions. The functions associated with a government may change over time, and uncommon functions may be associated with small governments, which have a small probability of selection.

2.3 Estimation

The composite estimation methodology is described here in a step-by-step manner, and various quantities and estimators are introduced along the way. Let k be an index used to refer to elements in the population (U) or sample (s). Variable y is measured in counts or dollars depending on the study variable. Y is a population total, and $\hat{Y}$ is an estimated total. Two important quantities are the previous CoG total for domain fd,

$$Y_{fd}^{\text{CoG}} = \sum_{k \in U_{fd}^{\text{CoG}}} y_k^{\text{CoG}},$$

and the raw (unweighted) sum of the sample data,

$$\hat{Y}_{fd}^{\text{RAW}} = \sum_{k \in s_{fd}} y_k.$$

The direct estimator equals the weighted sum,

$$\hat{Y}_{fd}^{\text{DIR}} = \sum_{k \in s_{fd}} w_k y_k,$$

where w_k is the sample weight for element k. Because $w_k \geq 1$ and $y_k \geq 0$, the direct estimator is always greater than or equal to the corresponding raw sum. The synthetic estimator equals

$$\hat{Y}_{fd}^{\text{SYN}} = \hat{Y}_f^{DB} K_{fd}^{\text{CoG}},$$

where $\hat{Y}_f^{DB}$ is the decision-based estimator of the state total (Shao *et al.*, 2014), and K_{fd}^{CoG} is the proportion of the function within the state from the most recent CoG,

$$K_{fd}^{\text{CoG}} = \frac{\sum_{k \in U_{fd}^{\text{CoG}}} y_k^{\text{CoG}}}{\sum_{k \in U_f^{\text{CoG}}} y_k^{\text{CoG}}}.$$

Next comes the preliminary composite estimator. It takes the form of a weighted average of the direct and synthetic estimators,

$$\hat{Y}_{fd}^{\text{COM1}} = \hat{\phi}_f \hat{Y}_{fd}^{\text{DIR}} + \left(1 - \hat{\phi}_f\right) \hat{Y}_{fd}^{\text{SYN}},$$

where

$$\hat{\phi}_f = 1 - \frac{\sum_d \hat{V}\left(\hat{Y}_{fd}^{\text{DIR}}\right)}{\sum_d \left(\hat{Y}_{fd}^{\text{DIR}} - \hat{Y}_{fd}^{\text{SYN}}\right)^2}$$

is a state-level James-Stein composite weight intended to allow for accurate estimation for the group of domains in state f as whole (Rao, 2003, p. 63). If $\hat{\phi}_f$ is initially estimated to be negative, then it is set to 0.5.

Several adjustments are applied to $\hat{Y}_{fd}^{\text{COM1}}$ to ensure certain quality control checks are met. The most important check is that the final estimate be greater than or equal to the raw sum. If $\hat{Y}_{fd}^{\text{COM1}}$ violates this check, a less-than-raw adjustment sets the final estimate equal to the direct estimate. As a simplification, the final composite estimator can be written as

$$\hat{Y}_{fd}^{\text{COM}} = \begin{cases} \hat{Y}_{fd}^{\text{COM1}}, & \hat{Y}_{fd}^{\text{COM1}} \geq \hat{Y}_{fd}^{\text{RAW}} \\ \hat{Y}_{fd}^{\text{DIR}}, & \hat{Y}_{fd}^{\text{COM1}} < \hat{Y}_{fd}^{\text{RAW}}. \end{cases}$$

3. Monte Carlo Simulation

3.1 Planning

A large Monte Carlo simulation is performed to study the design-based properties of the quantities and estimators described in section 2. To take advantage of the external validation offered by the CoG, we use public-use micro-data from the 2007 and 2012 CoG. All of the following figures and tables use data from those sources. Data from the 2007 CoG serve as the sampling frame, from which many independent samples are selected according to the current sample design. The samples are merged with the 2012 CoG data and then used to estimate the known 2012 totals, which are denoted Y_{fd}^{2012}. To mimic production as closely as possible, governments in 2007 that are known to disincorporate later are not removed from the sampling frame. The number of simulated samples, R, is set to 10,000 to obtain accuracy up to two decimal places when estimating proportions.

3.2 Performance Measures

Table 2 lists the performance measures that are calculated for the various estimators. These measures include absolute relative bias (ARB), coefficient of variation (CV), relative root mean squared error ($RRMSE$), and the proportion of less-than-raw occurrences (LTR). Global performance measures such as average $RRMSE$, denoted $\overline{RRMSE}$, are obtained by averaging over domains. The notation $est_{fd}^{(r)}$ stands for the estimate of Y_{fd}^{2012} for a particular estimator based on the r^{th} sample. On an added note, because public employment and payroll totals tend to be stable, these results arguably apply to estimators used during production in intercensal years.

Table 2: Performance Measures

Performance Measure	Formula
Expectation	$E_{fd} = \frac{1}{R}\sum_{r=1}^{R} est_{fd}^{(r)}$
Bias	$B_{fd} = E_{fd} - Y_{fd}^{2012}$
Absolute relative bias	$ARB_{fd} = \frac{\lvert B_{fd}\rvert}{Y_{fd}^{2012}}$
Variance	$V_{fd} = \frac{1}{R-1}\sum_{r=1}^{R}\left(est_{fd}^{(r)} - E_{fd}\right)^2$
Coefficient of variation	$CV_{fd} = \frac{\sqrt{V_{fd}}}{Y_{fd}^{2012}}$
Mean squared error	$MSE_{fd} = \frac{1}{R}\sum_{r=1}^{R}\left(est_{fd}^{(r)} - Y_{fd}^{2012}\right)^2$
Relative root mean squared error	$RRMSE_{fd} = \frac{\sqrt{MSE_{fd}}}{Y_{fd}^{2012}}$
Proportion of less-than-raw occurrences	$LTR_{fd} = \frac{1}{R}\sum_{r=1}^{R} 1\left(est_{fd}^{(r)} < \hat{Y}_{fd}^{\text{RAW}(r)}\right)$

3.3 Alternative Composite Estimators

Two alternative composite estimators are created for study purposes and are motivated by preliminary simulation results for the direct and synthetic estimators. The alternative preliminary composite estimator is given by

$$\hat{Y}_{fd}^{\text{ALTCOM1}} = \hat{\phi}_{fd}^{\text{ALT}} \hat{Y}_{fd}^{\text{DIR}} + \left(1 - \hat{\phi}_{fd}^{\text{ALT}}\right) \hat{Y}_{fd}^{\text{SYN}},$$

where

$$\hat{\phi}_{fd}^{\text{ALT}} = \frac{MSE_{fd}^{\text{SYN}}}{MSE_{fd}^{\text{SYN}} + MSE_{fd}^{\text{DIR}}}$$

is a near-optimal domain-specific weight based on the estimated mean squared errors of the direct and synthetic estimators (Rao, 2003, p. 58). There are several adjustments that are applied at the preliminary composite step, but as a simplification, we express the alternative preliminary composite estimator as the following to reflect just the less-than-raw adjustment,

$$\hat{Y}_{fd}^{\text{ALTCOM}} = \begin{cases} \hat{Y}_{fd}^{\text{ALTCOM1}}, & \hat{Y}_{fd}^{\text{ALTCOM1}} \geq \hat{Y}_{fd}^{\text{RAW}} \\ \hat{Y}_{fd}^{\text{DIR}}, & \hat{Y}_{fd}^{\text{ALTCOM1}} < \hat{Y}_{fd}^{\text{RAW}}. \end{cases}$$

Note that $\hat{\phi}_{fd}^{\text{ALT}}$ depends on the true totals Y_{fd}^{2012} through the estimated mean squared errors MSE^{DIR} and MSE^{SYN}. Consequently, $\hat{Y}_{fd}^{\text{ALTCOM1}}$ and $\hat{Y}_{fd}^{\text{ALTCOM}}$ are not real estimators for 2012, but they do provide a baseline for how accurate a composite estimator can be and further show the effect of the less-than-raw adjustment.

3.4 Simulation Results

Tables 3 and 4 display simulation results for *ftemp* and *ptemp*, which are representative of the full-time and part-time variables. The bias-variance tradeoff among the direct, synthetic, and preliminary composite estimators is reflected in the values of $\overline{ARB}$ and $\overline{CV}$. The direct estimator has small bias and large variance, whereas the synthetic estimator has relatively large bias and small variance. As a weighted average, the preliminary composite estimator balances the two.

Based on the values for $\overline{RRMSE}$ for all of the estimators, it appears easier to estimate totals accurately for *ftemp* than for *ptemp*. This makes sense because the full-time variables are more stable than the part-time variables. For *ptemp*, the value 45.52 percent for $\overline{RRMSE}$ for the preliminary composite estimator is very large. In fact, it is larger than the value 28.87 percent for $\overline{RRMSE}$ for the previous CoG total, which has no sampling variability. The less-than-raw adjustment that is applied to the preliminary composite estimator reduces $\overline{LTR}$ to 0 for both variables, but, in the case of *ftemp*, the adjustment increases $\overline{RRMSE}$ from 12.61 percent to 13.91 percent.

As expected, the alternative composite estimators from section 3.3 perform the best in terms of $\overline{RRMSE}$. This should give confidence in the simulation and the resulting estimates of variance and mean squared error. The increase in $\overline{RRMSE}$ caused by the less-than-raw adjustment is more pronounced for the alternative composite estimators. In the case of *ftemp*, $\overline{RRMSE}$ increases from 6.41 percent to 11.24 percent after the less-than-raw adjustment is applied.

Table 3: Monte Carlo Results for *ftemp*

Estimator	$\overline{ARB}\%$	$\overline{CV}\%$	$\overline{RRMSE}\%$	$\overline{LTR}\%$
Y_{fd}^{CoG}	14.94	0.00	14.94	12.31
$\hat{Y}_{fd}^{\text{RAW}}$	22.85	2.78	23.30	0.00
$\hat{Y}_{fd}^{\text{DIR}}$	0.16	17.85	17.85	0.00
$\hat{Y}_{fd}^{\text{SYN}}$	13.94	1.48	14.23	13.34
$\hat{Y}_{fd}^{\text{COM1}}$	6.15	9.18	12.61	7.04
$\hat{Y}_{fd}^{\text{COM}}$	5.85	10.68	13.91	0.00
$\hat{Y}_{fd}^{\text{ALTCOM1}}$	4.12	4.26	6.41	1.60
$\hat{Y}_{fd}^{\text{ALTCOM}}$	4.32	9.13	11.24	0.00

Data Source: U.S. Census Bureau, 2007 and 2012 Census of Governments: Employment

Table 4: Monte Carlo Results for *ptemp*

Estimator	$\overline{ARB}\%$	$\overline{CV}\%$	$\overline{RRMSE}\%$	$\overline{LTR}\%$
Y_{fd}^{CoG}	28.87	0.00	28.87	10.31
$\hat{Y}_{fd}^{\text{RAW}}$	45.97	4.86	45.71	0.00
$\hat{Y}_{fd}^{\text{DIR}}$	0.74	75.13	75.14	0.00
$\hat{Y}_{fd}^{\text{SYN}}$	29.66	7.85	32.72	10.47
$\hat{Y}_{fd}^{\text{COM1}}$	14.94	39.13	45.52	5.26
$\hat{Y}_{fd}^{\text{COM}}$	16.07	37.76	45.36	0.00
$\hat{Y}_{fd}^{\text{ALTCOM1}}$	13.33	11.63	19.02	2.09
$\hat{Y}_{fd}^{\text{ALTCOM}}$	15.95	12.47	22.13	0.00

Data Source: U.S. Census Bureau, 2007 and 2012 Census of Governments: Employment

It was mentioned in section 2.3 that the composite weight $\hat{\phi}_f$ can be estimated to be negative initially. Figure 1 plots the proportion of occurrences of negative weights for *ftemp* versus average sampling fraction. These proportions and averages are with respect to the 10,000 samples from the Monte Carlo simulation, and the sampling fraction refers to the elements in the 2012 data. Each data point represents a single state, and of the 49 states, 40 have proportions very close to one. It is reasonable to think that the larger the sampling fraction, the better the estimate of $\hat{\phi}_f$ and the smaller the proportion of negative composite weights. However, the plot suggests no such relationship. Florida, Alaska, Delaware, and Nevada, which have different sampling fractions, all have proportions exactly equal to 0. The plot versus the actual sample size is similar.

If $\hat{\phi}_f$ is initially estimated to be negative, then it is set to 0.5. In the absence of information regarding the direct and synthetic estimates, the value 0.5 is an intuitive choice as it puts equal weight on the two. Another option is to set the weight to 0, as is done with the positive-part James-Stein estimator (Lehmann and Casella, 1998, p. 275). For ASPEP, this would mean letting the final estimate be the synthetic estimate, which is stable but has the undesirable property of being less than the raw sum about ten percent of the time. All in all, the composite methodology used during ASPEP's 2009 sample design appears to lack a smooth estimator of the composite weight.

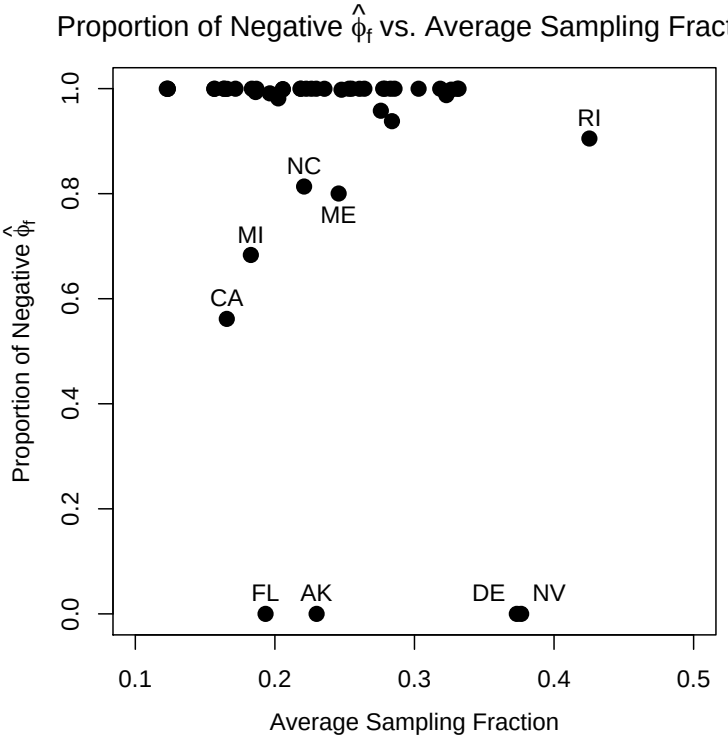

Figure 1: Plot of proportion of occurrences of negative $\hat{\phi}_f$ for *ftemp* versus average sampling fraction. Each data point represents a single state. The proportions and averages are with respect to the 10,000 samples from the Monte Carlo simulation, and the sampling fraction refers to elements in the 2012 data. Data Source: U.S. Census Bureau, 2007 and 2012 Census of Governments: Employment

4. Hierarchical Bayesian Models

4.1 Motivation

The composite estimator used during ASPEP's 2009 sample design is based on an implicit area-level model. In area-level models (Rao, 2003, ch. 5), the direct survey estimates are modeled, and any covariates are at the area level. One advantage of area-level models is that they are valid for general sample designs and usually result in design-consistent estimators. Area-level models for ASPEP have a lot of potential because there exist very good covariates in the form of totals from the most recent CoG. As an illustration, Figure 2 plots the strong linear relationship between 2007 and 2012 CoG totals on the log scale for *ftemp* in Illinois.

In a unit-level model, on the other hand, values of individual units are modeled, and covariates can be at the unit or area level. A type of unit-level model called a nested-error regression model has been used successfully during ASPEP's 2014 sample design (Tran and Dumbacher, 2014). In this approach, the model is fitted using sample data without the sampling weights and then used to make predictions for the out-of-sample units. However, if approximate design consistency of the resulting estimators is desired, then greater care needs to be taken to account for the sample design.

The goal of this part of the research is to explore explicit area-level models that try to capture the spirit of the composite methodology and address areas for improvement in a unified and generalizable way. The hierarchical Bayesian method (Rao, 2003, ch. 10),

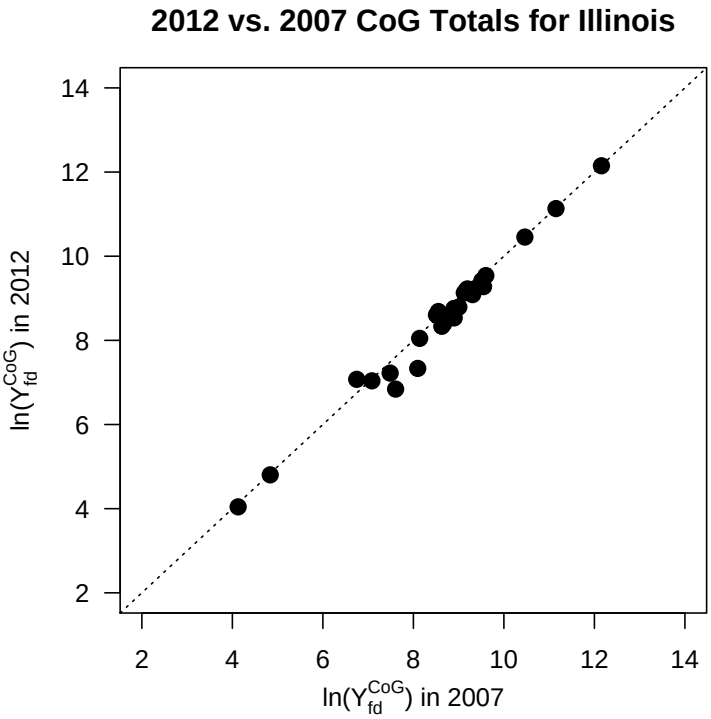

Figure 2: Plot of 2012 CoG totals versus 2007 CoG totals on the log scale for *ftemp* in Illinois. Each dot represents a function. The line of equality has been added for reference. Data Source: U.S. Census Bureau, 2007 and 2012 Census of Governments: Employment

which is new in application to ASPEP, offers a convenient setting in which to fit and evaluate such models. 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 in the data. This type of modeling involves modeling regression coefficients to tie the various levels in the data together. Two good resources on the subject are Gelman and Hill (2007) and Gelman *et al.* (2014).

4.2 Notation

This section introduces notation for the area-level models considered in this research. First of all, to make modeling assumptions more justified and to reduce the scale of the data, we work with log-transformed direct estimates and parameters. Let $\hat{\theta}_{fd} = ln(\hat{Y}_{fd}^{\text{DIR}})$ and $\theta_{fd} = ln(Y_{fd})$ be the log-transformed direct estimate and true total, respectively, for domain fd, where $f = 1, \ldots, F$ and $d = 1, \ldots, D_f$. Denote by $\hat{\boldsymbol{\theta}}_f = (\hat{\theta}_{f1}, \ldots, \hat{\theta}_{fD_f})^T$ and $\boldsymbol{\theta_f} = (\theta_{f1}, \ldots, \theta_{fD_f})^T$ the vectors of transformed direct survey estimates and parameters for state f, respectively. Also, let $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ denote similar quantities for all domains nationwide.

Denote by $\psi_{fd} = V(\hat{\theta}_{fd})$ the design-based variance of $\hat{\theta}_{fd}$ and by $\boldsymbol{\Psi}_f = V(\hat{\boldsymbol{\theta}}_f)$ the design-based variance-covariance matrix of $\hat{\boldsymbol{\theta}}_f$. The parameters ψ_{fd} and $\boldsymbol{\Psi}_f$ are assumed known but are actually estimated using the 10,000 direct estimates from the Monte Carlo simulation in section 3.

4.3 Fay-Herriot Model

The most well-known area-level model for continuous data is the Fay-Herriot (FH) model, which was first used to estimate per capita income in 1970 for places having fewer than 1,000 people according to the 1970 Census of Population of Housing (Fay and Herriot, 1979). In the context of complex survey data, the FH model can be written in terms of a sampling model with "sampling errors" and a linking model with "model errors." The errors are assumed to be independent and normally distributed.

The following FH model is fit separately for each state and serves as a building block for more complicated models. For state f and functions $d = 1, \ldots, D_f$,

$$\begin{aligned}
\hat{\theta}_{fd}|\theta_{fd} &\overset{ind}{\sim} N(\theta_{fd}, \psi_{fd}) \\
\theta_{fd}|\boldsymbol{\beta}_f, \sigma_f^2 &\overset{ind}{\sim} N(\boldsymbol{x}_{fd}^T\boldsymbol{\beta}_f, \sigma_f^2) \\
\boldsymbol{\beta}_f &\sim N_k(\mathbf{0}, 100\boldsymbol{I}_k) \\
\sigma_f &\sim U(0, 100),
\end{aligned}$$

where $\boldsymbol{\beta}_f$ is a $k \times 1$ vector of regression coefficients, σ_f^2 is the hierarchical variance, $\boldsymbol{x}_{fd} = (1, ln(Y_{fd}^{\text{CoG}}))^T$ is the vector of covariates that includes the transformed value of the previous CoG total, and $\boldsymbol{I}_k$ is the $k \times k$ identity matrix. Weakly informative prior distributions are used because flat, improper priors cannot be specified using the model fitting software and are reasonable choices to represent a lack of information relative to the data input. Also, in hierarchical models, Gelman (2006) recommends a uniform prior for σ_f over the commonly used inverse-Gamma distribution.

4.4 Fay-Herriot Model with Correlated Sampling Errors

As pointed out by Xie, Raghunathan, and Lepkowski (2007), one limitation of the standard FH model is that it assumes independent sampling errors among the domains. Because the sample design for ASPEP is complicated and the fact that the domains cut across clusters of elements, the sampling errors in the first-level model are probably correlated. To accommodate this, a multivariate extension of the FH model is considered that includes a general variance-covariance matrix for the transformed direct survey estimates. This is the same model considered by Datta and Lahiri (1995). Formally, the model is given by the following. For state f and functions $d = 1, \ldots, D_f$,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_f\,|\boldsymbol{\theta}_f &\sim N_{D_f}(\boldsymbol{\theta}_f, \boldsymbol{\Psi}_f) \\
\theta_{fd}|\boldsymbol{\beta}_f, \sigma_f^2 &\overset{ind}{\sim} N(\boldsymbol{x}_{fd}^T\boldsymbol{\beta}_f, \sigma_f^2) \\
\boldsymbol{\beta}_f &\sim N_k(\mathbf{0}, 100\boldsymbol{I}_k) \\
\sigma_f &\sim U(0, 100),
\end{aligned}$$

where $\boldsymbol{\Psi}_f$ is an estimate of the designed-based variance-covariance matrix $V(\hat{\boldsymbol{\theta}}_f)$. The covariance estimates come from the Monte Carlo simulation, and the formula is similar to the variance formula given in Table 2. For state f and functions d and d', the covariance of

$\hat{\theta}_{fd}$ and $\hat{\theta}_{fd'}$ is estimated by

$$Cov\left(\hat{\theta}_{fd}, \hat{\theta}_{fd'}\right) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}_{fd}^{(r)} - \overline{\hat{\theta}_{fd}^{(\cdot)}}\right)\left(\hat{\theta}_{fd'}^{(r)} - \overline{\hat{\theta}_{fd'}^{(\cdot)}}\right),$$

where

$$\overline{\hat{\theta}_{fd}^{(\cdot)}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{fd}^{(r)} \qquad \text{and} \qquad \overline{\hat{\theta}_{fd'}^{(\cdot)}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{fd'}^{(r)}.$$

4.5 Multilevel Fay-Herriot Models

The previous models are fit separately for each state. It may be advantageous to borrow strength across states and jointly model the transformed direct survey estimates nationwide. To this end, multilevel models are considered that tie the state-level models together through modeling of the $\boldsymbol{\beta}_f$. The formulation below allows for a general variance-covariance matrix $\boldsymbol{\Psi}_f$, but this multilevel model can be fit to either the case of independent errors or that of correlated sampling errors. For state $f = 1, \ldots, F$ and functions $d = 1, \ldots, D_f$,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_f \,|\boldsymbol{\theta}_f &\overset{ind}{\sim} N_{D_f}(\boldsymbol{\theta}_f, \boldsymbol{\Psi}_f) \\
\theta_{fd}|\boldsymbol{\beta}_f, \sigma_f^2 &\overset{ind}{\sim} N(\boldsymbol{x}_{fd}^T\boldsymbol{\beta}_f, \sigma_f^2) \\
\boldsymbol{\beta}_f \,|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\overset{iid}{\sim} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\mu} &\sim N_k(\mathbf{0}, 100\boldsymbol{I}_k) \\
\boldsymbol{\Sigma} &\sim Inv\text{-}Wishart(k+1, \boldsymbol{I}_k) \\
\sigma_f &\overset{iid}{\sim} U(0, 100).
\end{aligned}$$

4.6 Summary of Models

Table 5 summarizes the four models considered in this research.

Table 5: Summary of Models

Acronym	Description
FH	Fay-Herriot
$FHcor$	Fay-Herriot with correlated sampling errors
MFH	Multilevel Fay-Herriot
$MFHcor$	Multilevel Fay-Herriot with correlated sampling errors

5. Model Fitting and Evaluation

5.1 Model Fitting

The models are fit in R (R Core Team, 2012) using the package *rjags*. JAGS (Plummer, 2012) stands for Just Another Gibbs Sampler and is the Unix equivalent of the popular

WinBUGS software (Lunn *et al.*, 2000), which allows one to construct Bayesian models and simulate draws from posterior distributions using Markov chain Monte Carlo (MCMC) methods (Gilks, Richardson, and Spiegelhalter, 1996). The model syntax of JAGS is very similar to that of WinBUGS. The R packages *coda* and *R2WinBUGS* are used to assess convergence of the Markov chains and to analyze the posterior output.

For the FH and $FHcor$ models, we use five chains, dispersed initial values, a burn-in of 50,000 iterations followed by 100,000, and a thinning rate of ten. This results in a posterior sample size of 50,000 for each quantity of interest. For the MFH and $MFHcor$ models, we reduce the number of iterations by a factor of five to deal with computer memory issues. Standard MCMC diagnostics indicate good mixing of the chains, low autocorrelation, and potential scale reduction factors close to one.

5.2 Model Evaluation

The models are evaluated in terms of how well replicated data agree with the observed data and how well the model estimates agree with the true 2012 totals. Under squared error loss, the Bayes estimate of a quantity of interest is the posterior mean. We denote the Bayes estimate of the true 2012 total $Y_{fd}^{2012} = exp(\theta_{fd})$ by $\hat{Y}_{fd}^{\text{BAYES}}$.

One way to evaluate and diagnose model fit in the Bayesian framework is through posterior predictive checks, which are described in Gelman *et al.* (2014). A popular discrepancy measure used in such checks is the omnibus χ^2 measure (Gelman, Meng, and Stern, 1996). This measure is given by

$$\chi^2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{f=1}^{F} \sum_{d=1}^{D_f} \frac{\left[\hat{\theta}_{fd} - E(\hat{\theta}_{fd}|\boldsymbol{\theta})\right]^2}{V(\hat{\theta}_{fd}|\boldsymbol{\theta})} = \sum_{f=1}^{F} \sum_{d=1}^{D_f} \frac{\left(\hat{\theta}_{fd} - \theta_{fd}\right)^2}{\psi_{fd}},$$

where the expectation E and variance V are with respect to the model. Variations of this discrepancy measure are considered by limiting the set of functions over which the inner summation is taken. This can give insight into what functions are difficult to estimate. For example, we restrict attention to the elementary and secondary education functions (012 and 112), which make up a large proportion of public employment and payroll, and the public utility functions (091, 092, and 093), which, historically, have been difficult to estimate accurately. The posterior predictive p-value (ppp) is given by

$$ppp = P\left(\chi^2(\hat{\boldsymbol{\theta}}^{\text{rep}}, \boldsymbol{\theta}) > \chi^2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \;\middle|\; \hat{\boldsymbol{\theta}}\right),$$

where the probability is with respect to the joint posterior distribution of replicated data and the parameters, $\pi(\hat{\boldsymbol{\theta}}^{\text{rep}}, \boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$. In general, values of ppp close to 0 and 1 indicate lack of fit, while values close to 0.5 indicate good fit.

To evaluate estimation accuracy, the square root of the average squared distance between the model estimates and true totals Y_{fd}^{2012} is calculated. This distance is denoted Δ and is given by

$$\Delta = \sqrt{\frac{1}{M} \sum_{f=1}^{F} \sum_{d=1}^{D_f} \left(\hat{Y}_{fd}^{\text{BAYES}} - Y_{fd}^{2012}\right)^2},$$

where

$$M = \sum_{f=1}^{F} D_f$$

is the number of domains to which the model under consideration can be applied. As with the posterior predictive checks, we can restrict attention to the elementary and secondary education and public utility functions.

CVs are also useful for comparisons. The average CV, which is denoted $\overline{CV}$, and the maximum CV, which is denoted CV_{max}, are also calculated. The average and maximum are taken over all M domains.

6. Modeling Results

6.1 Results

Tables 6 and 7 display modeling results for *ftemp* and *ptemp* after applying the four models to a single sample. Results for the composite estimator $\hat{Y}_{fd}^{\text{COM}}$ are added for comparison. The subscripts ED and PU stand for elementary and secondary education and public utilities, respectively.

Table 6: Modeling Results and Comparisons for *ftemp*

Model	χ^2	χ^2_{ED}	χ^2_{PU}	Δ	Δ_{ED}	Δ_{PU}	$\overline{CV}\%$	$CV_{max}\%$
FH	0.16	0.54	0.39	621.97	1,712.95	742.38	5.97	55.08
$FHcor$	0.14	0.34	0.46	609.52	1,673.44	734.40	5.76	54.92
MFH	0.15	0.53	0.38	623.77	1,726.38	733.77	5.91	51.39
$MFHcor$	0.12	0.37	0.44	611.27	1,688.81	724.29	5.70	52.42
$\hat{Y}_{fd}^{\text{COM}}$	–	–	–	768.58	2,298.02	552.58	5.48	91.38

Data Source: U.S. Census Bureau, 2007 and 2012 Census of Governments: Employment

Table 7: Modeling Results and Comparisons for *ptemp*

Model	χ^2	χ^2_{ED}	χ^2_{PU}	Δ	Δ_{ED}	Δ_{PU}	$\overline{CV}\%$	$CV_{max}\%$
FH	0.03	0.51	0.15	1,285.18	1,538.92	70.27	21.09	172.04
$FHcor$	0.03	0.52	0.13	1,263.40	1,445.19	65.88	20.17	167.00
MFH	0.03	0.51	0.14	1,280.02	1,499.13	62.91	20.43	171.66
$MFHcor$	0.02	0.54	0.13	1,259.03	1,409.02	58.89	19.58	156.12
$\hat{Y}_{fd}^{\text{COM}}$	–	–	–	1,330.69	1,821.82	154.86	31.74	496.11

Data Source: U.S. Census Bureau, 2007 and 2012 Census of Governments: Employment

6.2 Discussion

For *ftemp*, the *ppp* values for the χ^2 measures do not indicate severe model misfit. Considering the Δ and CV measures, the models with correlated sampling errors, $FHcor$ and $MFHcor$, perform the best, but the simpler $FHcor$ has an edge in terms of estimation accuracy. The composite estimator does perform well in estimating the totals for the public utility functions, and its value for $\overline{CV}$ is smaller than that of any of the models. However, its values of 768.58 for Δ and 91.38 percent for CV_{max} are large.

For *ptemp*, which is the more volatile of the two study variables considered here, the *ppp* values for the omnibus χ^2 measure are closer to 0. However, the benefit of the mod-

eling approach over composite estimation is reflected in the other measures. The most complicated model, $MFHcor$, which has both correlated sampling errors and multilevel structure, performs the best according to all Δ and CV measures. $FHcor$ performs well too, just like for *ftemp*, but multilevel modeling for *ptemp* seems to result in an appreciable improvement in estimation accuracy. Composite estimation performs worse than the other models for *ptemp*.

7. Future Research

The models considered here show promise, but they can serve as building blocks for more complicated models. One interesting extension assumes either the sampling errors or model errors follow a t-distribution, the heavy-tailed nature of which provides robustness against outliers. Working in a Bayesian framework, Bell and Huang (2006) applied these robust models to data from the Small Area Income and Poverty Estimates program and found that assuming a t-distribution for the model (sampling) errors pushes the final estimate closer to the direct (synthetic) estimate. This has interesting implications for ASPEP because the direct estimate is always greater than the corresponding raw sum.

Future research could also involve investigating other discrepancy measures. The posterior predictive p-values from the omnibus χ^2 measures do not indicate severe model misfit for *ftemp*, so it would be useful to find measures that check for other practical features of the data. One criticism of posterior predictive checks is that they make double use of the data. The data are used to train the prior into a posterior and again when being compared against replicated data from the fitted model. Cross-validated posterior predictive checks are an alternative (Larsen and Lu, 2007), and entire states in the multilevel models seem like natural cross-validation "folds."

Lastly, although the prior distributions currently assumed should be uninformative relative to the data input, it is good practice to examine the sensitivity of the results to the choice of prior. For example, we could consider different upper limits for the uniform prior on σ_f.

REFERENCES

Bell, W.R. and Huang, E.T. (2006). Using the *t*-Distribution to Deal with Outliers in Small Area Estimation. *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health*. Statistics Canada.

Datta, G.S. and Lahiri, P. (1995). Robust Hierarchical Bayes Estimation of Small Area Characteristics in the Presence of Covariates and Outliers. *Journal of Multivariate Analysis*, *54*(2), 310–328.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, *74*(366), 269–277.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D.B. (2014). *Bayesian Data Analysis*, Third Edition. Boca Raton, FL: Chapman & Hall.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gelman, A., Meng, X.L., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (Eds.). (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: Chapman & Hall.

Gonzalez, M.E. (1973). Use and Evaluation of Synthetic Estimates. *Proceedings of the Social Statistics Section*, American Statistical Association, 33–36.

Larsen, M.D. and Lu, L. (2007). Comment: Bayesian Checking of the Second Level of Hierarchical Models: Cross-Validated Posterior Predictive Checks Using Discrepancy Measures. *Statistical Science*, *22*(3), 359–362.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, Second Edition. New York, NY: Springer-Verlag.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – A Bayesian Modeling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing*, *10*, 325–337.

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

Plummer, M. (2012). JAGS Version 3.3.0 User Manual. Available at <mcmc-jags.sourceforge.net>

R Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at <www.R-project.org>

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons, Inc.

Shao, J., Slud, E., Cheng, Y., Wang, S., and Hogue, C. (2014). Theoretical and Empirical Properties of Model Assisted Decision-Based Regression Estimators. *Survey Methodology*, *40*(1), 81–104.

Tran, B. and Cheng, Y. (2011). Application of Small Area Estimation for Annual Survey of Employment and Payroll. *2011 Proceedings of the American Statistical Association, Survey Research Methods Section*. Alexandria, VA: American Statistical Association, 3659–3673.

Tran, B. and Dumbacher, B. (2014). An Evaluation of Different Small Area Estimators for the Annual Survey of Public Employment and Payroll. *2014 Proceedings of the American Statistical Association, Survey Research Methods Section*. Alexandria, VA: American Statistical Association, 153–163.

U.S. Census Bureau. (2006). Government Finance and Employment Classification Manual. Available at <www2.census.gov/govs/pubs/classification/2006_classification_manual.pdf>

U.S. Census Bureau. (2014). 2013 Annual Survey of Public Employment & Payroll Methodology. Available at <www2.census.gov/govs/apes/2013_methodology.pdf>

Xie, D., Raghunathan, T.E., and Lepkowski, J.M. (2007). Estimation of the Proportion of Overweight Individuals in Small Areas–A Robust Extension of the Fay-Herriot Model. *Statistics in Medicine*, *26*(13), 2699–2715.

Appendix A: Functions

Code	Function
001	Air transportation
005	Corrections
012	Elementary and secondary education – Instructional
112	Elementary and secondary education – Other
016	Higher education – Other
018	Higher education – Instructional
023	Financial administration
024	Fire protection – Firefighters
124	Fire protection – Other
025	Judicial and legal
029	Other government administration
032	Health
040	Hospitals
044	Highways
050	Housing and community development
052	Libraries
059	Natural resources
061	Parks and recreation
062	Police protection – Persons with power of arrest
162	Police protection – Other
079	Public welfare
080	Sewerage
081	Solid waste management
087	Water transport and terminals
089	All other and unallocable
091	Water supply
092	Electric power
093	Gas supply
094	Transit

Source: U.S. Census Bureau

Additional information about the classification of government employment activities can be found in U.S. Census Bureau (2006).