Application of Small Area Estimation for Annual Survey of Employment and Payroll

Bac Tran, Yang Cheng
Governments Division
U.S. Census Bureau[1], Washington, D.C. 20233-0001

Abstract: Annual Survey of Employment and Payroll estimates the number of federal, state, and local government employees and their gross payrolls. In the past two years, we developed the Decision-based method to estimate the survey total. In this paper, we discuss some small area challenges when we estimate the survey total at the functional level of government units such as airport, public welfare, hospitals, etc. First, we introduce the synthetic estimation and modified direct estimators. Then, we modified the composite estimation as a weighted average between modified direct estimation and synthetic estimation. Finally, we evaluate these methods by using the 2007 Census of Governments: Employment Component.

Key Words: Decision-based Estimation, Modified Direct Estimator, Synthetic Estimation, Composite Estimation

1. Introduction

The Annual Survey of Public Employment and Payroll (ASPEP) produces statistics on the number of federal, state, and local government employees and their gross payrolls. For more information on the survey, please see Website for ASPEP http://www.census.gov/govs/apes/. ASPEP provides current estimates for full-time and part-time state and local government employment and payroll by government function (i.e., elementary and secondary education, higher education, police protection, fire protection, financial administration, judicial and legal, etc.). ASPEP covers all states and local governments in the United States, which include counties, cities, townships, special districts, and school districts. The first three types of government are referred to as general-purpose governments, because they generally provide multiple government activities. Activities are coded as function codes. School districts cover only education functions. Special districts usually provide only one function, but can provide two or three functions. ASPEP is the only source of public employment data by program function and selected job category. Data on employment include number of full-time and part-time employees, gross pay, and hours paid for part-time employees. Reported data are for the government's pay period that includes March 12. Data collection begins in March and continues for about seven months.

There are 89,526 state and local government units in our universe. In 2009, after exploring possible cut-off sample methods for ASPEP, we developed a new modified cut-off sample method based on the current stratified probability proportional-to-size (PPS) sample design. This method reduced the sample size, which saved resources, improved the precision of the estimates, reduced respondent burden, and improved data quality. The modified cut-off sample method was applied in two stages. We first

[1] *Disclaimer: This report is released to inform interested parties of research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.*

selected a state-by-governmental type stratified PPS sample. The PPS sample was based on total payroll, which was the sum of full-time pay and part-time pay, from the Employment portion of the 2007 Census of Government. In the second stage, we constructed a cut-off point to distinguish small and large government units in the stratum. Lastly, we sub-sampled the stratum with small-size government units.

ASPEP was designed to estimate survey totals of key variables: full-time employment, full-time payroll, part-time employment, part-time payroll, part-time hours, full-time equivalent employment, total payroll, and total employment. Cheng et al. (2009) proposed a method, Decision-based, to improve the precision of estimates and reduce the mean square error of weighted survey total estimates. Basically, the Decision-based method combined the strata to improve the models by testing the equality of the slopes of regression models from different strata. In Cheng et al. (2009), the hypothesis test was carried out in two steps. First, a test was performed of the null hypothesis that the slopes were identical. If the p-value was less than 0.05, the null hypothesis would be rejected to conclude that the regression lines were significantly different. In this case, there was no reason to compare the intercepts. If the p-value was greater than 0.05, the null hypothesis of equality of slopes could not be rejected, but intercepts could be compared. If the regression lines for the two substrata were not found to be significantly different, then a single line was estimated from the combined substrata. The Decision-based estimates provided a fundamental base to improve the reliability of the indirect small area estimation.

The ASPEP's sampled units were stratified by state and government types. However, it was required to estimate the variables of interest at the state and functional code level, which contained up to 30 categories for each government unit. This naturally brought the small area challenges, because we did not have any control on the sample size at the state and function code level. For example, the sample size for the state of Maryland was 48. But, there were only 3 samples units airport activity, labeled as function code of 001. In the worst case, we have zero sample for some specific function codes. If there were missing data in some specific function for a government unit, these missing data could be structural zeros. We define that structural zeros to be cells in which observations are impossible. Table 1 shows each government unit in a state may have different functions. Table 2 lists all government function codes.

Table 1: There are structural zeros in the government unit (marked as N/A)

	GOVERNMENT UNITS							
FUNCTION	**1**	**2**	**3**	**4**	**5**	**…**	**N-1**	**N**
Airport	✓	N/A	N/A	N/A	N/A	…	✓	N/A
Correction	✓	✓	N/A	✓	✓	…	✓	✓
Elementary/Second	✓	✓	✓	✓	N/A	…	N/A	✓
Financial	N/A	✓	✓	✓	✓	…	✓	✓
FireFighters	✓	✓	✓	✓	✓	…	✓	✓
…	…	…	…	…	…	…	…	…
Fire	✓	✓	✓	✓	✓	…	✓	✓
Police	✓	N/A	✓	✓	✓	…	✓	✓

Table 2: Function codes in the Annual Survey of Public Employment and Payroll

ItemCode	Meaning
000	Totals for Government
001	Airports
002	Space Research & Technology (Federal)
005	Correction
006	National Defense and International Relations (Federal)
012	Elementary and Secondary - Instruction
112	Elementary and Secondary - Other Total
014	Postal Service (Federal)
016	Higher Education - Other
018	Higher Education - Instructional
021	Other Education (state)
022	Social Insurance Administration (state)
023	Financial Administration
024	Firefighters
124	Fire - Other
025	Judicial and Legal
029	Other Government Administration
032	Health
040	Hospitals
044	Streets & Highways
050	Housing & Community Development (Local)
052	Local Libraries
059	Natural Resources
061	Parks & Recreations
062	Police Protection - Officers
162	Police - Other
079	Welfare
080	Sewerage
081	Solid Waste Management
087	Water Transport & Terminals
089	Other & Unallocable
090	Liquor Stores (state)
091	Water Supply
092	Electric Power
093	Gas Supply
094	Transit

What is Small Area? Small Area is a small geographic area within a larger geographic area or a small demographic group within a larger demographic group. The sample size in the domain of interest is too small to use a standard estimator. Most small area estimation methods borrow strength from related or similar areas using auxiliary data. There is growing demand from the public for reliable small area statistics. At the design stage, we don't consider attaining precision at the state and function code level. However, we have to handle this challenge at the estimation stage.

Let g represent state and f represent function code level. We want to estimate the total of employees or payroll information at the state by function level:

$$Y_{gf} = \sum_{i \in U_{gf}} Y_{gfi}$$

where U is the universe of function codes in all states, and U_{gf} is the universe of function code f, state g. Thus, U_{gf} is subset of U, that is, $U_{gf} \subset U$. The sample size for function code f, n_f, is less than or equal to the sample size n, that is, $n_f \leq n$. The domain of sample for function code level f of state g is the intersection of the sample domain of state g and the universe of function code f and state g, $S_{gf} = S_g \cap U_{gf}$.

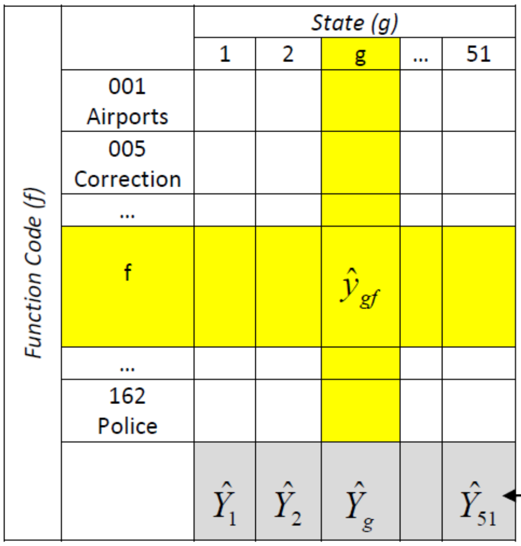
In some cases, the changes in Employment statistics are relatively stable. Therefore, a linear regression is suitable for some state by government type cells as done prior to Fiscal Year (FY) 2009. However, due to small sample sizes and poor fits on many cells, a small area estimation method (SAE) is more appropriate. SAE is only applied on PPS sample. For certainties, the direct estimate was used. Information on Births and Non-Activity (B&N) units is not available at the sampling stage. Therefore, we sample B&N separately from the PPS and Certainties sample.

Figure 1 briefly shows how we estimated the variable of interest in each cell of state by function code table. We applied the design-based direct estimator (Horvitz-Thompson), and the synthetic estimator in each cell. The direct estimator has high variability due to the small sizes. On the other hand the synthetic estimator reduces the variability but introduces some bias. Therefore, we introduce the composite estimator, which is a weighted average of those two estimators. We also modified the direct estimator (modified direct) from borrowing strength from similar cells to smooth the direct estimator. We will go through each of our estimators in detail in subsequent sections.

2. Methodology

In this section, we discuss how to estimate Y_{gf} for a given state g and function code f. Here, Y represents the survey total of key variables: full-time employment, full-time payroll, part-time employment, part-time payroll, part-time hours, full-time equivalent employment, total payroll, and total employment. We describe all the estimators used in our estimation process: Direct (Horvitz-Thompson), Decision-based, Synthetic, Composite, Modified Direct, and the Composite estimator.

Figure 1: Cross-Tabulation of State by Function



		State (g)				
Function Code (f)		1	2	g	...	51
	001 Airports					
	005 Correction					
	...					
	f			$\hat{y}_{gf}$		
	...					
	162 Police					
		$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_g$		$\hat{Y}_{51}$

2009 ASPEP regress on 2007 Census (decision-based)

2.1 Direct estimator (Horvitz-Thompson)

A general design-based direct estimator for the total is:

$$\hat{t}_{y,gf} = \sum_{i \in S} w_{i,gf}\, y_{i,gf} \,. \qquad (1)$$

where the weight, $w_i = \dfrac{1}{\pi_i}$, and π_i is the inclusion probability for unit i in state g and function code f. In this paper, we also denote $\hat{t}_{y,gf}$ as $\hat{Y}_{gf}^{HT}$.

2.2 Decision-based estimator

The Decision-based (DB) method helps to estimate the synthetic in each cell by providing a stable state total as a reliable estimator in a large area covering all small areas, states by function code level. DB was a process of testing the possibility of combining the strata. This strengthened statistical models for the area of estimation. The state total was estimated by a single stratum weighted regression (GREG) estimator specified as follows:

$$\hat{t}_{y,GREG} = \hat{t}_{y,\pi} + \hat{b}(t_x - \hat{t}_{x,\pi}) \qquad (2)$$

where $t_x = \sum_{i \in U} x_i$, $\hat{t}_{x,\pi} = \sum_{i \in S} \dfrac{x_i}{\pi_i}$, $\hat{t}_{y,\pi} = \sum_{i \in S} \dfrac{y_i}{\pi_i}$, $\hat{b} = \dfrac{\sum_{i \in S} (x_i - \bar{x})(y_i - \bar{y}) / \pi_i}{\sum_{i \in S} (x_i - \bar{x})^2 / \pi_i}$, π_i is the inclusion probability, and x_i is the auxiliary data from the Employment portion of the Census of Governments for government unit i.

The slope $\hat{b}$ was obtained by the Decision-based (DB) process proposed by Cheng et al. (2009). The DB method improved the precision of estimates and reduced the mean square error of weighted survey total estimates. The idea was to test the equality of linear regression lines to determine whether we can combine data in different substrata. The null hypothesis $H_0 : b_1 = b_2$, that is, the equality of the frame population regression slopes for two substrata. In large samples, $\hat{b}$ is approximately normally distributed, $\hat{b} \sim N(b, \Sigma\ )$. Under the null hypothesis, with two sub-strata U_1, U_2 from samples S_1, S_2 of sizes n_1, and n_2, we have $\hat{b}_1 - \hat{b}_2 \sim N(0, \Sigma_{1,2})$ where $\hat{b}_1 \sim N(b, \sum_1\ )$, $\hat{b}_2 \sim N(b, \sum_1\ )$, and $\sum_{1,2} = \sum_1 + \sum_2$. Therefore, the test statistic is

$$(\hat{b}_1 - \hat{b}_2)\Sigma_{1,2}^{-1}(\hat{b}_1 - \hat{b}_2) \sim \chi_1^2 \tag{3}$$

Our research showed that it was unnecessary to do the hypothesis for the intercept equality because our data analysis showed that we never rejected the null hypothesis of equality of intercepts when we could not reject the null hypothesis of equality of slopes. This is reasonable because the 2007 payroll could be 0 essentially only if the 2002 payroll was.

We will discuss the variance estimator for $\hat{b}$ in Section 3. The critical value for a test based on (3) was obtained from a chi-squared distribution with 1 degree of freedom. The test was performed with a significance level of $\alpha = 0.05$. If we could not reject the null hypothesis, then the slopes estimated in sub-strata S_1 and S_2 were accepted as the same, and the Decision-based estimator was equal to the GREG estimator for the union of two sample sets, that is, for $S = S_1 \cup S_2$. Otherwise, the Decision-based estimator would be the sum of two separate GREG estimators of stratum totals, that is,

$$\hat{t}_{y,DB} = \begin{cases} \hat{t}_{y,greg} & \text{if } \ \mathrm{H}_0 \ \text{ is accepted} \\ \sum_{h=1}^{2} \hat{t}_{y,greg}^{h} & \text{if } \ \mathrm{H}_0 \ \text{ is rejected.} \end{cases} \tag{4}$$

where $\hat{t}_{y,greg}$ denotes the GREG estimator from the combined stratum S, while $\hat{t}_{y,greg}^{h}$ denotes the GREG estimator from substratum h from sample S_h. DB produced 51 (50 states and Washington D.C.) totals for each key variable.

2.3 Synthetic estimation

Synthetic estimation assumes that small areas have the same characteristics as large areas, and there is a valid unbiased estimate for large areas. There are many advantages of synthetic estimation. They are accurate aggregated estimates, simple and intuitive, applied to all sample designs, and borrow strength from similar small areas. Synthtic estimation can even provide estimates for areas with no sample from the sample survey, and it does not need a study model.

The general idea for synthetic estimation is that if we have a reliable unbiased estimate for a large area and this large area covers many small areas, then we can use this estimate to produce an estimate for a small area. The key element for calculating the synthetic estimation for a small area (state by function code level) is to estimate the proportion of that small area of interest within the large state area. This estimate for the small area is known as the synthetic estimate.

The synthetic estimator for function code *f* of state *g* is:

$$\hat{Y}^{S}_{gf} = \frac{x_{gf}}{\sum_{f} x_{gf}} \hat{t}^{DB}_{g} \qquad (5)$$

where x_{gf} is auxiliary information which is obtained from Employment portion of Census of Government and the state total, $\hat{t}^{DB}_{g}$ is obtained by the Decision-based from equation (4).

2.4 Composite estimator

To balance the potential bias of the synthetic estimator, $\hat{Y}^{S}_{gf}$, against the instability of the design-based direct estimator, $\hat{Y}^{HT}_{gf}$, we introduce a composite estimator as a weighted average of these two estimators. Thus, the composite estimate was applied on the PPS sample for each state by function code cell. Generally, it has the form:

$$\hat{Y}^{C}_{gf} = \phi_{g} \hat{Y}^{HT}_{gf} + (1-\phi_{g}) \hat{Y}^{S}_{gf}. \qquad (6)$$

where $\hat{\phi}_{g} = 1 - \frac{\sum \mathrm{var}(\hat{y}^{HT}_{gf})}{\sum (\hat{y}^{S}_{gf} - \hat{y}^{HT}_{gf})^{2}}$ (Purcell & Kish, 1979). In some cases, we observed negative $\hat{\phi}$. To fix this problem, we applied the method which was introduced by Lahiri and Pramanik (2010). They suggested using average mean square errors (AMSE) instead of MSE to compute $\hat{\phi}$.

2.5 Modified direct estimator

We replaced the direct $\hat{Y}^{HT}_{gf}$ in (6) by a modified direct estimate (MD), $\hat{Y}^{MD}_{gf}$, due to instability of the design-based direct estimate caused by small sizes. The modified direct estimator from Rao's Small Area Estimation (2003) is given as:

$$\hat{Y}^{MD}_{gf} = \hat{Y}^{HT}_{gf\pi} + \hat{b}_{f} (X_{gf} - \hat{X}^{HT}_{gf\pi}) \qquad (7)$$

where

$$\hat{Y}^{HT}_{gf,\pi} = \sum_{i\in S_{gf}} \frac{y_{gf,i}}{\pi_{g,i}},\; X_{gf} = \sum_{i\in U_{gf}} x_{gf,i},\; \hat{X}^{HT}_{gf,\pi} = \sum_{i\in Sgf} \frac{x_{gf,i}}{\pi_{g,i}}, \text{ and}$$

$$\hat{b}_f = \frac{\sum_{g\in G, i\in Sgf} (x_{gf,i} - \bar{x}_f)(y_{gf,i} - \bar{y}_f)/\pi_{g,i}}{\sum_{g\in G, i\in Sgf} (x_{gf,i} - \bar{x}_f)^2/\pi_{g,i}}$$

Since the modified direct estimators use data from outside the domain, we can see that the MD method is smoothed by borrowing strength across the state. The estimator $\hat{Y}^{MD}_{gf}$ is approximately unbiased as the overall sample size increases, even if the domain sample size is still small. The modified direct estimator (7) is performed under some conditions which allowed producing a reliable $\hat{b}_f$, for example, goodness of fit R^2, slopes, and the sample sizes.

One good example for the MD estimator is the case of missing reported data for Louisiana and Mississippi due to hurricane Katrina. Modified direct estimates used information outside the domain of interest, and the regression coefficient $\hat{b}_f$ was the same across the state of Louisiana. The MD estimator is a regression estimator, approximately unbiased. Finally, the modified direct estimator is a calibration estimator if written as an expansion direct form by minimizing the chi-square distance subject to the constraints with calibration property.

2.6 Modified composite estimator

With MD estimator available, we can modify the composite estimator as:

$$\hat{Y}^{C}_{gf} = \phi_g \hat{Y}^{MD}_{gf} + \left(1-\phi_g\right)\hat{Y}^{S}_{gf} \qquad (8)$$

We can re-write the MD estimator as:

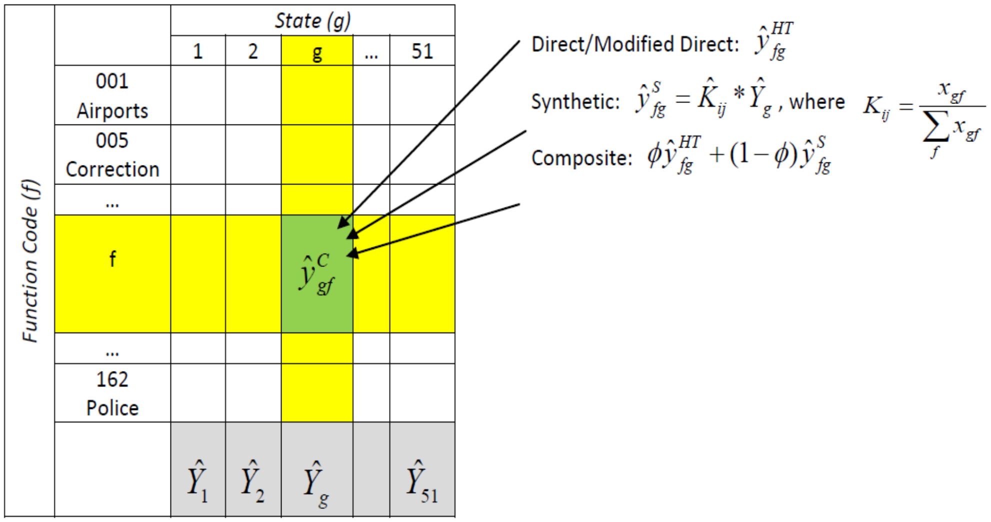
$$\hat{Y}^{MD}_{gf} = X_{gf} * \hat{b}_f + \sum_{j\in S_{gf}} w_j e_j \qquad (9)$$

where

$$e_j = y_j - X_j * \hat{b}_f$$

The first term $X_{gf}\hat{b}_f$ is the synthetic regression estimator and the second term, $\sum_{j\in S_{gf}} w_j e_j$ approximately corrects the bias of the synthetic estimator. Figure 2 shows all the estimators we discuss in this paper.

Figure 2: Cross-Tabulation of State by Function Estimators in Each Cell


State (g)
1
2
g
...
51
Function Code (f)
001 Airports
005 Correction
...
f
$\hat{y}_{gf}^{C}$
...
162 Police
$\hat{Y}_1$
$\hat{Y}_2$
$\hat{Y}_g$
$\hat{Y}_{51}$
Direct/Modified Direct: $\hat{y}_{fg}^{HT}$
Synthetic: $\hat{y}_{fg}^{S} = \hat{K}_{ij} * \hat{Y}_g$, where $K_{ij} = \frac{x_{gf}}{\sum_f x_{gf}}$
Composite: $\phi\hat{y}_{fg}^{HT} + (1-\phi)\hat{y}_{fg}^{S}$

3. Variance Estimation

Due to the complexity of the two-stage sampling design with the cut-off technique, we calculated the approximate variance (AV) of the composite estimator. AV is estimated on the non sub-sampling sample. Besides, there are B&N units, which are very small and contribute a small amount in the survey total. We assume the variance on B&N is ignorable.

The coefficient of variance, CV, is estimated by $\frac{\sqrt{Var(\hat{y})}}{\hat{y}}$, where $\hat{y}$ is the composite estimated on PPS, certainties, and B&N.

The government units were sampled by state and government type. However, the variance is required for the cell of state and function code, which we don't know the size in advance. Therefore, in order to estimate the variance for the cell state by function code, we treated each combination D = (state, function code) as a domain in the sample, which is identified by the indicator I_D = 1 if the unit belongs to D, and 0 otherwise. We used Taylor series method for the variance estimation in which the variation among different units of the same function code was taken into account. The domain total is:

$\hat{Y}_D = \sum_g \sum_f \sum_j I_D \, w_{hfj} \, y_{fgj}$, where g = state, f = function code, and j= unit jth.

The estimated variance is:

$$\hat{V}(\hat{Y}_D) = \sum_g \hat{V}_g(\hat{Y}_D), where$$

If $n_g > 1$,

$$\hat{V}_g(\hat{Y}_D) = \frac{n_g(1-f_g)}{n_g-1} \sum_{i=1}^{n_g} (z_{gi.} - \bar{z}_{g..})^{\,2}$$

where
$z_{gf.} = \sum_{j=1}^{m_{gf}} I_D w_{gfj} y_{gfj}, \; \bar{z}_{g..} = (\sum_{i=1}^{n_h} z_{gf.})/n_g$

If $n_g = 1$ then

$$\hat{V}_g(\hat{Y}_D) = \begin{cases} \text{missing} & if\ n_{g'} = 1\ for\ g' = 1,2,..,G \\ 0 & if\ n_{g'} > 1\ for\ some\ 1 \le g' \le G \end{cases}$$

Notes: States, DC and Hawaii had CV =0 because they are census.

This variance can be estimated using proc surveymeans in SAS software.

4. Empirical Bayes Estimate vs. the Composite Estimate

This section shows a connection between Empirical Bayes (EB) and the composite estimate in a simple case where the design variance D is assumed constant over domains. Also, with the assumption of normality of the data, there is a similarity between the composite estimation method and the empirical Bayes method. The small difference appears in the composite weight and the shrinkage coefficient (see below). We also included some Empirical Bayes (EB) estimates results when we conducted the data analysis on state by function code level in the Employment and Payroll information.

With two-level model (Lahiri, 2006):

Level 1: $\hat{y}_{gf}^D \,|\theta_{gf} \sim N(\theta_{gf}, D)$

Level 2: $\theta_{gf} \sim N(x_{gf}^T \beta, A)$

the empirical estimate of the variable of interest is

$$\hat{\theta}_{gf}^{EB} = B\hat{y}_{gf}^D + (1-B)\hat{x}_{gf}^T \hat{\beta} \qquad (10)$$

where *A* and *β* are unknown, *m* is the number of small areas, and the shrinkage coefficient is:

$$\hat{B} = 1 - \frac{(m-1)D}{\sum_{i=1}^{m} (\hat{y}_{gf}^D - x_{gf}^T \beta)^2}.$$

It is a very slight different between $\hat{B}$ and $\hat{\phi}$, where $\hat{\phi} = 1 - \frac{mD}{\sum_{i=1}^{m} (\hat{y}_{gf}^D - x_{gf}^T \beta)^2}$.

We also present the comparison between the empirical Bayes and the composite estimates in some states (see Section 5)**.**

5. Results

The composite estimator was used to estimate the survey totals in each cell (state by function) of the ASPEP. As mentioned earlier, the composite estimator is the weighted average of the two estimators: the design-based and the synthetic. The composite balances out the instability of the unbiased due to small sample sizes with the synthetic quantity. The weight ϕ pulls the estimate to the design unbiased estimate when it has enough data, and towards the synthetic estimate when there is insufficient sample size in the small area (Rao, 2003).

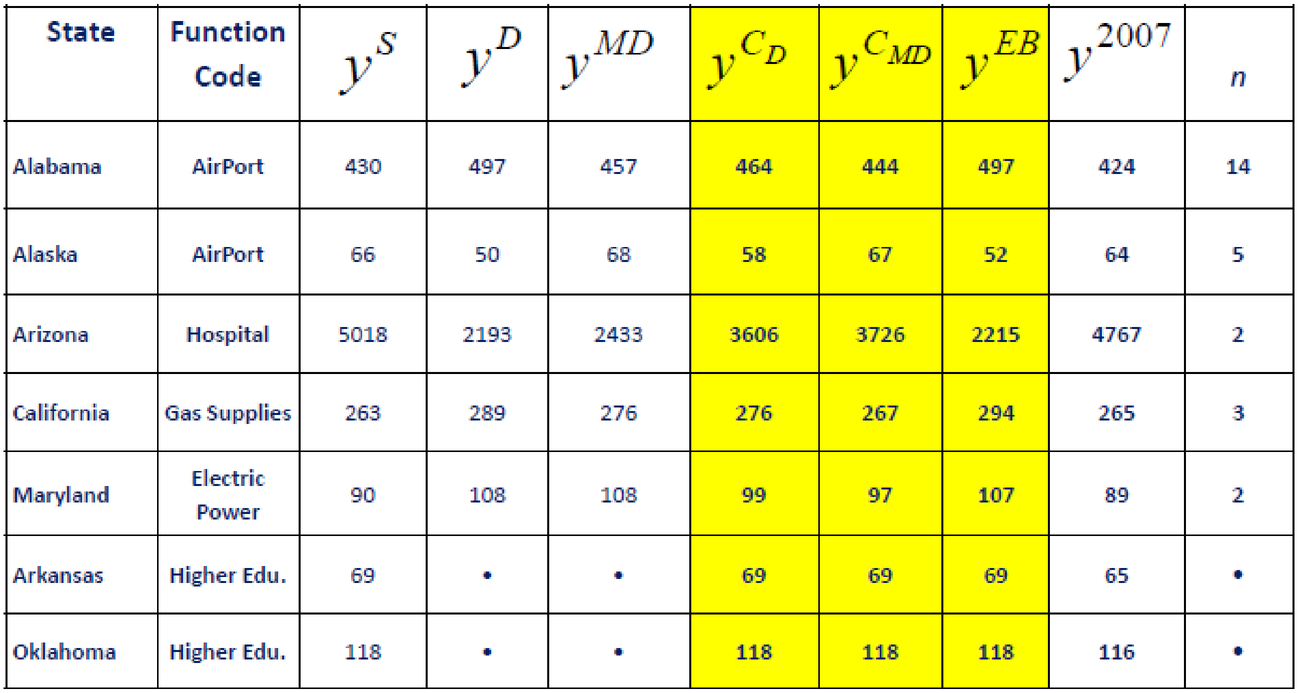
By applying the methods described in Section 2, we created Table 3 which is a typical illustration of our data analysis. Those methods included a combination of Decision-based estimation and an application of a SAE method. Table 3 is for the variable, Full Time Equivalent Employment, in several randomly selected states. The 2007 data (census data) is included in the Table 3 to see the changes of the variable overtime from different estimators. It is not used to evaluate different estimators. However, for some stable variable like Full Time Employees, 2007 census data is useful to see the performance of the estimators. The conclusions are as follows:

- When there were no observed sampled units, we used the synthetic estimate where the design-based direct estimates were not present. For example, there were no samples units in higher education in Arkansas or Oklahoma, we obtained a reasonable synthetic estimate.
- The synthetic estimates were stable in small size areas where the design-unbiased estimates were very volatile.
- The modified direct estimates were closer to the 2007 census values.
- When the sample sizes were big enough, all the estimators performed well and they were close to each other.
- The composite using the modified direct estimator was close to the 2007 Census values most often.

Figure 3 shows the comparison among the composite estimate, synthetic estimate, design-based direct estimate (Horvitz-Thompson), and the 2007 data (census) for the variable, Full Time Employees, in Alabama for all functions from the most recent Census of Governments (2007). Figure 4 is a snapshot from Figure 3 which focuses on function codes 080, 081, 089, 091, and 092. Figure 4 shows the performance of the synthetic and the composite over the design-based estimate. Figures 3 & 4 show that when the sample sizes are relatively small the synthetic and the composite estimates outperformed the design-based estimates.

Note: Code 080 and 091 are sewerage and water supplies which are problematic because respondents cannot separate the data for the two variables. Code 089 is problematic because it is a catch-all "All other" variable, which tends to be volatile.

Table 3: Comparison of Different Estimators in Various Sample Sizes



State	Function Code	y^{S}	y^{D}	y^{MD}	y^{C_D}	$y^{C_{MD}}$	y^{EB}	y^{2007}	*n*
Alabama	AirPort	430	497	457	464	444	497	424	14
Alaska	AirPort	66	50	68	58	67	52	64	5
Arizona	Hospital	5018	2193	2433	3606	3726	2215	4767	2
California	Gas Supplies	263	289	276	276	267	294	265	3
Maryland	Electric Power	90	108	108	99	97	107	89	2
Arkansas	Higher Edu.	69	•	•	69	69	69	65	•
Oklahoma	Higher Edu.	118	•	•	118	118	118	116	•

Figure 3: Comparison the Estimates Composite, Synthetic, and Horvitz-Thompson for the variable Full Time Employees in Alabama (all functions)

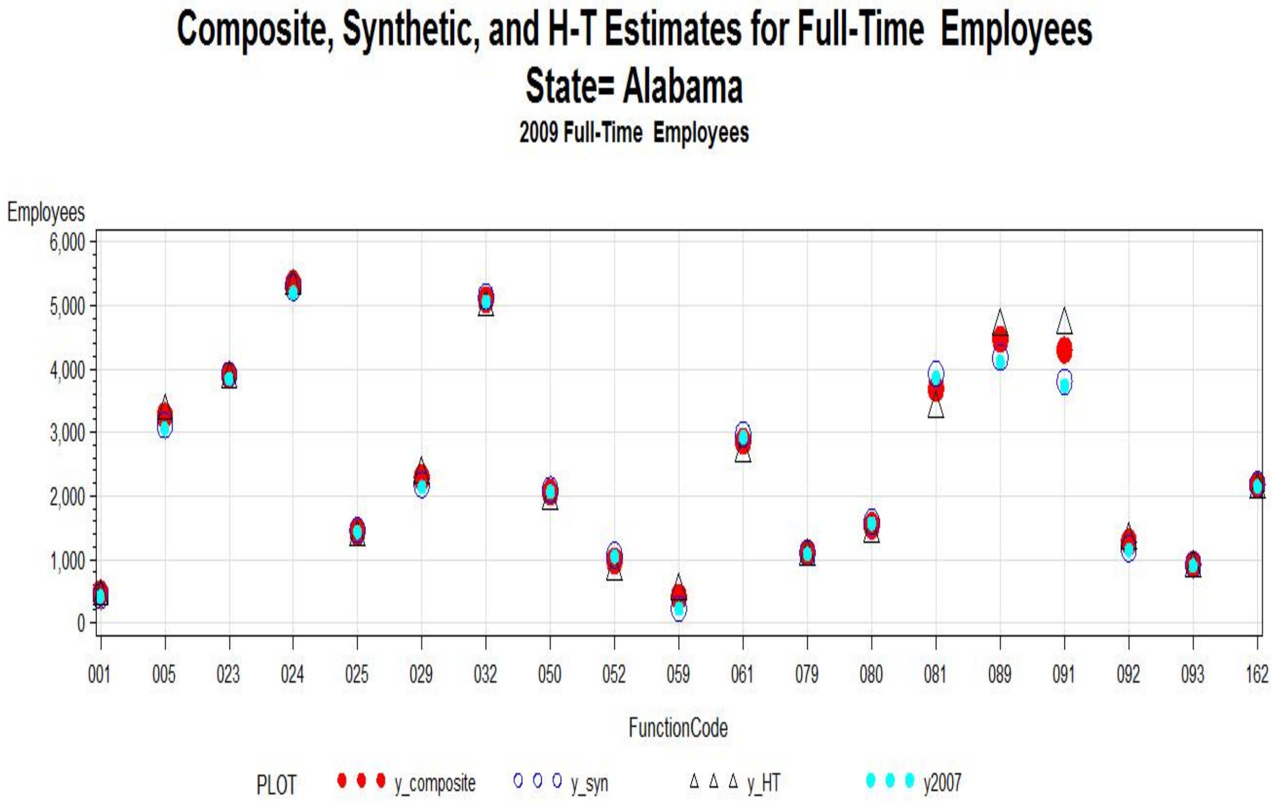
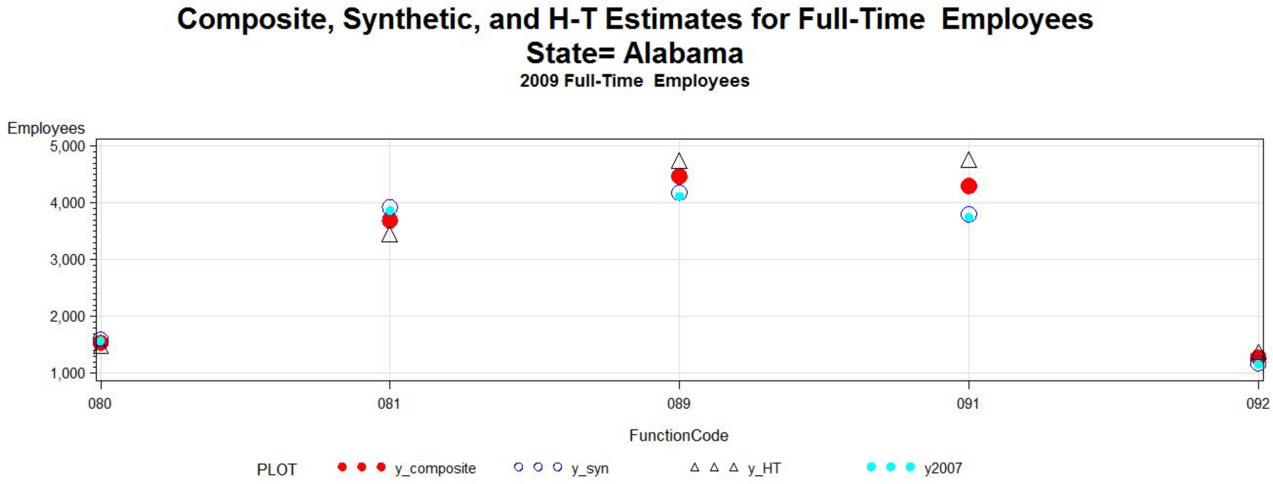


Figure 4: Comparison the Estimates Composite, Synthetic, and Horvitz-Thompson for the variable Full Time Employees in Alabama



6. Conclusions

Bias of the synthetic estimator is the biggest disadvantage for synthetic estimation. Departures from the assumption may lead to large biases. Empirical studies have mixed results on the accuracy of synthetic estimators. The bias may not be estimated from the data.

The variance estimator for the complicated composite estimator derived from a Decision-based method needs separate research which will be presented in a future paper.

This paper presents two applications: Decision-based and Small Area Estimation methods. They were applied to the estimation of Annual Survey of Public Employment and Payroll. SAE provides the composite estimate which smoothes the design unbiased estimators in small areas by introducing the synthetic term. The synthetic estimate is more reliable when derived from the Decision-based estimates. This property cannot be obtained from a simple regression synthetic.

With these two methods arecombined, we obtained better estimates than those of using direct estimators or with linear regression where the linear relationship is weak or even does not exist.

7. Future Research

We have some outstanding issues which need further research. We need to develop a simple and good variance estimator formula for the composite estimator other than a resampling method. Regarding the weight, $\hat{\phi}_g$, in the composite estimation method, we replace $\hat{\phi}_g$ = 0.5 when it was negative. Lahiri and Pramanik (2010) extended a method from Gonzalez & Waksberg (1973), which used Average Design-based Mean Squared Error (AMSE) to stabilized the $\hat{\phi}_g$. We will apply this method in our production in the future. We will also explore in more detail the application of the Empirical Bayes method with an alternative assumption other than normality. Finally, we will apply this method for other surveys in the Governments, like the Annual Finance Survey (AFS).

Acknowledgements

We would like to thank Dr. Partha Lahiri from the University of Maryland at College Park and the Center for Statistical Research and Methodology of the Census Bureau for contributing ideas about average mean square errors and bridge between composite estimation and Empirical Bayes method. Also, we are indebted to our reviewers, Lisa Blumerman and Carma Hogue from the Governments Division of the Census Bureau. We also thank Dr. William Bell from the Associate Director of Research and Methodology of the Census Bureau for his valuable suggestions, which improved the original research.

References

Barth, J., Cheng, Y., Hogue, C. (2009). Reducing the Public Employment Survey Sample Size, *JSM Proceedings.*

Cheng, Y., Corcoran, C., Barth, J., Hogue, C. (2009). An Estimation Procedure for the New Public Employment Survey, *JSM Proceedings.*

Cheng, Y., Slud, E., Hogue, C. (2010). Variance Estimation for Decision-based Estimators with Application to the Annual Survey of Public Employment and, *JSM Proceedings.*

Cochran, W.G. (1977), Sampling Techniques. Third Edition. New York: John Wiley & Sons, Inc.
Deville, J-C. and Sarndal, C-E. (1992), Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, Volume 87, Number 418, 376-382

Lahiri, P. and Pramanik, S. (2010), Discussion of "Estimating Random Effects via Adjustment for Density Maximization" by C. Morris and R. Tang, Statistical Science, Volume 26, Number 2, 291-295

Purcell, N.J., and Kish, L. (1979), Estimates for Small Domain, Biometrics, 35, 365-384

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

Sarndal, C.-E., Swensson, B., and Wretman, J. (1992), Model Assisted Survey Sampling, Springer-Verlag.