Evaluating Calibration Estimators for the Annual Survey of Local Government Finances

Elizabeth L. Love[1], Joseph J. Barth, Bac Tran[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
Elizabeth.L.Love@census.gov, Joseph.J.Barth@census.gov, Bac.Tran@census.gov

Abstract

The Governments Division of the U.S. Census Bureau conducts the Annual Survey of Local Government Finances (ALFIN). The ALFIN provides statistics about the financial activities of state and local governments across the country. We currently use calibration to estimate these finance statistics. Calibration methods adjust sampling weights so that the adjusted weight totals agree with reliable known totals, e.g., census totals (or census counts) obtained from the Census of Governments. In previous cycles of the ALFIN, survey analysts used decision-based estimation, a technique that performs hypothesis tests that allow combining strata when possible to reduce the variance and improve the accuracy of survey estimates. In this evaluation, we develop a design-based Monte Carlo simulation experiment in which we draw repeated samples from the 2007 Census of Governments data using the ALFIN sample design. We compute the decision-based, calibration, and Horvitz-Thompson estimates that use the generated sample and the 2002 Census of Governments data as auxiliary information. We then compare mean squared errors of these estimators.

Keywords: Governments Unit; Annual Survey of State and Local Government Finances; Calibration

1. Introduction

The U.S. Census Bureau conducts the Annual Survey of Local Government Finances (ALFIN) that collects data about the financial activities of local governments across the United States. Estimates based on the data collected by this sample survey estimate the revenues, expenditures, debts, and assets of local governments. We publish these local statistics along with their corresponding state level aggregates from the Annual Survey of State Government Finances. Data from both of these surveys assist in the development of the government component of the Gross Domestic Product estimates, the allocation of some federal grant funds, and public policy research.

2. Sample Design

This section describes the ALFIN sample that was designed in 2009. The Finance Component of the 2007 Census of Governments (CoG) provided the auxiliary information used for the size variable and identifying certainty units on the frame. The

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.

sampling frame contained the units surveyed in 2007, updated with births (newly created governments) and deaths (governments that closed) from the 2008 survey cycle. The sample can be divided into multiple parts based on the financial activity of each government. What follows is a description of each activity-based grouping and how that grouping was selected.

The sample can be broken down into two parts: certainty units and non-certainty units. Criteria for certainty units included all school districts, large-population general-purpose governments (cities, counties, townships) and special districts with sizeable amounts of long-term debt, revenue, and expenditure reported in the 2007 CoG. Certainty units accounted for 55% of the units in the sample.

The remainder of the units falls into the non-certainty part of the sample. The non-certainty units can be divided into two sub-groups. The first group contains births and non-activities (governments that did not report any debts or expenditures or reported values of zero for both). The sample is updated with births and non-activity units annually. Deaths are removed annually as well. We used simple random systematic sampling to select which units to add to the sample. These units made up less than one percent of the total sample size.

We selected the second part of the non-certainty units in two phases. The first phase was probability proportional-to-size (πps) sampling and the second was a modified cutoff sampling technique. In the first phase, we sampled from four of the five types of local governments: counties, municipalities, townships, and special districts. Each government type in a state defined a stratum. After stratifying the units, we used πps sampling where the size variable was the maximum of total expenditure and long-term debt in 2002 to determine which units should initially be in the sample. The second phase of sampling is used to reduce the number of non-contributing sub-counties and special districts in sample.

Modified cutoff sampling was our method for determining how many small units should be in the final sample. Government units in each stratum in the first-phase sample were divided into a large cutoff stratum and small cutoff stratum using the cumulative square root of the frequency method (Corcoran and Cheng, 2010).

3. Data

This section will describe the data collected from the ALFIN, the statistics published from the survey, and the data we chose to focus on in this evaluation.

Local governments across the country provide the data collected from the ALFIN. Each financial activity is coded by an item code. There are more than two hundred item codes on the survey. These item codes are grouped into one of four major categories: revenues, expenditures, assets, or debts. The statistics we publish from the ALFIN data are categorized by these four categories.

The downloadable and viewable files are the two statistical products we provide annually using ALFIN data. Estimates for the total of each item code at the local level of government by state as well as the nation are available for public use in the downloadable file. The viewable file is a table that contains aggregates of the item code totals in the four major categories with some of the more important detailed items. These statistics are given by state and are published online in a nested table. For example, an estimate of how much revenue was collected from each tax item code is listed individually in the downloadable file. The viewable file contains a total tax aggregate and is the sum of all the revenue collected from the taxes listed in the downloadable file (see Table 1 for an excerpt of a viewable file).

Table 1. 2011 National Local Estimates of Tax Revenue

Description	National Local Estimate
Taxes	578,177,391
Property	429,086,267
Sales and gross receipts	93,078,804
General sales	65,430,782
Selective sales	27,648,022
Motor fuel	1,344,338
Alcoholic beverage	511,719
Tobacco products	403,210
Public utilities	14,056,299
Other selective sales	11,332,456
Individual income	25,628,794
Corporate income	7,163,771
Motor vehicle license	1,662,382
Other taxes	21,557,373

Source: U.S. Census Bureau, 2011 Annual Surveys of State and Local Government Finances

The data products from the ALFIN generate many statistics. For simplicity, we decided to concentrate on estimating total revenue using the non-certainty units for each state. Other data exclusions include removing Hawaii and Washington, D.C because they only contain certainty units and data that are published along with the ALFIN data products but come from other surveys conducted by the Governments Division or data about intergovernmental revenue for local governments that are reported by state government units.

4. Estimation Methods

Survey researchers experimented with a variety of estimation approaches for the ALFIN since the 2009 survey cycle. This section will explain the statistics we estimate and provide technical information about the estimation methods that are being compared in this evaluation.

4.1 Estimation Goal

Consider the units from the Finance component of the Census of Governments as a finite population $U = \{1, \dots i, \dots N\}$. Let the value of the i^{th} unit for the revenue item code c in state k be represented as y_{kic}. Then total revenue in state k is given by:

$$t_{kc} = \sum_{i \in U} y_{kic}. \qquad (1)$$

Estimates of t_{kc} are found in the downloadable file. In order to create the viewable file, the estimates of those item code totals, $\hat{t}_{kc}$, are aggregated over one or more item codes to estimate the statistics on the viewable file. In this evaluation, our goal is to estimate total revenue for the non-certainty units in each state. The sum of all the revenue item codes in a state is its total revenue denoted as t_k. Then the total of interest, t_k, is given by:

$$t_k = \sum_{\forall c} t_{kc.} \tag{2}$$

In this evaluation, we are estimating total state revenue in 2007 for non-certainty units only.

The rest of this section describes two estimation methodologies we employed to provide survey estimates for the ALFIN. The first method used was decision-based estimation and was first put into production for the 2009 survey cycle. Two years later, we replaced decision-based estimation with calibration. We will explain the theory and how we applied it to ALFIN data for each method.

4.2 Decision-Based Estimation

The decision-based (DB) method provides a reliable estimate of a large area that covers all small areas of interest. In this evaluation, the large area is the state total revenue and the small area is the state total by item code. DB was a process of testing the possibility of combining the strata in order to get a better estimate of the total. This method strengthened the statistical models for the area of estimation. The state total for revenue was estimated by a weighted regression (GREG) estimator specified as follows:

$$\hat{t}_{ky}^{GREG} = \hat{t}_{ky,\pi} + \hat{b}_k(t_{kx} - \hat{t}_{kx,\pi}) \tag{3}$$

where $t_{kx} = \sum_{i\in U} x_{ki}$, $\hat{t}_{sx,\pi} = \sum_{i\epsilon s} \frac{x_{ki}}{\pi_{ki}}$,
$\hat{t}_{ky,\pi} = \sum_{i\epsilon s} \frac{y_{ki}}{\pi_{ki}}$, and

$$\hat{b}_k = \frac{\sum_{i\epsilon s}(x_{ki} - \bar{x}_k)(y_{ki} - \bar{y}_k)\big/\pi_{ki}}{\sum_{i\epsilon s}\ (x_{ki} - \bar{x}_k)^2\big/\pi_{ki}}.$$

where π_{ki} is the inclusion probability, x_{ki} is the value of the auxiliary variable, and y_{ki} is the value of the variable of interest from the sample data for unit *i* in state *k*.

The slope $\hat{b}_k$ was obtained by the DB process proposed by Cheng et al. (2009). In that paper, 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 was $H_0\colon b_{k1} = b_{k2}$, that is, the equality of the frame population regression slopes for two substrata. In large samples, $\hat{b}_k$ is approximately normally distributed, $\hat{b}_k \sim N(b, \Sigma)$. Under the null hypothesis, with two substrata U_1, U_2 (large and small) from samples S_1, S_2 of sizes n_1, n_2, we have $\hat{b}_{k1} - \hat{b}_{k2} \sim N(0, \Sigma_{1,2})$ where $\hat{b}_{k1} \sim N(b, \Sigma_1)$, $\hat{b}_{k2} \sim N(b, \Sigma_2)$, and $\Sigma_{1,2} = \Sigma_1 + \Sigma_2$. Therefore, the test statistic is

$$(\hat{b}_{k1} - \hat{b}_{k2})\Sigma_{1,2}^{-1}(\hat{b}_{k1} - \hat{b}_{k2}) \sim \chi^2_{1.} \qquad (4)$$

Prior research showed that it was unnecessary to test 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.

The critical value for a test based on (4) was obtained from a chi-squared distribution with 1 degree of freedom. The test was performed with a significance level of α = 0.05. If we could not reject the null hypothesis, then the slopes estimated in substrata S_1 and S_2 were accepted as the same, and the DB 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,

$$\hat{t}_k^{DB} = \begin{cases} \hat{t}_{ky}^{GREG}, & if\ H_0\ is\ accepted \\ \hat{t}_{ky1}^{GREG} + \hat{t}_{ky2}^{GREG} & if\ H_0\ is\ not\ accepted \end{cases} \qquad (5)$$

where $\hat{t}_{ky}^{GREG}$ denotes the GREG estimator from the combined stratum S, while $\hat{t}_{ky1}^{GREG}$ and$\hat{t}_{ky2}^{GREG}$ denote the GREG estimator from substratum 1 and 2 from sample S_1 and S_2 respectively. DB produced state totals for revenue.

4.3 Calibration Estimation

Calibration methods consist of adjusting the sample design weights so that survey estimates of totals agree with known population totals which could be obtained from external sources. Calibration estimators use auxiliary data to adjust the sampling weights with respect to a set of constraints called calibration equations.

Suppose there is a finite population $U = \{1, \dots, i, \dots, N\}$. Let a probability sample $s\ (s \subseteq U)$ be drawn with a given sampling design and assume that the inclusion probabilities $\pi_i = \Pr(i \in s)$ and the joint inclusion probabilities $\pi_{ij} = \Pr(i\ \&\ j \in s)$ are always positive. These assumptions become more important when estimating the variance using the method proposed by Deville and Särndal, (1992). Let y_i be the value of the variable of interest, y, for the i^{th} population element in U. Let x_i be the value of the auxiliary variable, x, for the i^{th} population element, x_i can contain many variables but this evaluation uses a single variable.

Suppose we observe (y_i, x_i) for $i \in s$ and that the population total, $t_x = \sum_U x_i$, is known. The goal is to find a set of weights, w_i, by adjusting the sample design weights, $d_i = \frac{1}{\pi_i}$, so that the constraint below is

$$\hat{t}_x = \sum_s w_i x_i. \qquad (6)$$

There are many estimators that satisfy this condition (see Deville and Särndal, 1992). Deville, Särndal, Sautory (1993) suggested choosing adjusted weights that meet (6) and are close to the survey weight. They called this class of estimators calibration estimators. There are many ways to define the function used to specify the distance between the design weights and the calibrated weights. How close the calibrated weights are to the survey weights can be measured by a distance function, $\boldsymbol{g}$. Linear distance functions can produce calibrated weights with undesirable properties like falling below one, or worse, being negative. To avoid this problem, we use SUDAAN, developed by Research Triangle Institute (RTI), a statistical software that uses nonlinear calibration weighting. Through repeated linearization, SUDAAN finds a $\boldsymbol{g}$ such that

$$\sum_{U} \boldsymbol{x}_i = \sum_{S} d_i \alpha(\boldsymbol{g}^T \boldsymbol{x}_i) \boldsymbol{x}_i \qquad (7)$$

where

$$\alpha(\boldsymbol{g}^T \boldsymbol{x}_i) = \frac{l(u-c) + u(c-l)\exp(A\boldsymbol{g}^T \boldsymbol{x}_i)}{(u-c) + (c-l)\exp(A\boldsymbol{g}^T \boldsymbol{x}_i)} \; and \; A = \frac{u-l}{(u-c)(c-l)}.$$

The l, u, and c terms are user-defined where l specifies the lower bound, u is the upper bound, and c is a centering parameter (Kott, 2011). Equation (7) is the general formula and is written for a vector of auxiliary variables but for this evaluation we used a single variable.

In this evaluation, c is one because we are not adjusting for nonresponse or undercoverage. We let the lower bound be zero and the upper bound be infinity (technically it is 10^{20}) which are the default settings in SUDAAN. Also, we used a no-intercept model because it suits the data well and improves model fit.

We find total revenue for each government unit by summing all of the revenue item codes for that unit as follows:

$$x_i = \sum_{\forall c} x_{ic}. \qquad (8)$$

We calibrate using known state revenue totals from the 2007 Cog-F as follows:

$$t_k = \sum_{i \epsilon U, i \epsilon k} x_i. \qquad (9)$$

Now let's denote the sample design weights as $\{d_i\}$. We want to find a set of calibrated weights denoted as $\{w_i\}$ by adjusting $\{d_i\}$ such that the linear distance between the $\{d_i\}$ and the $\{w_i\}$ is minimized and satisfies the following constraint:

$$t_k = \sum_{i\epsilon s, i\epsilon k} w_i x_i. \tag{10}$$

In the production environment, the x_{ic} terms come from auxiliary data provided by the most recent CoG. In this evaluation, the 2002 Finance component of the CoG supplied the auxiliary data, x_{ic}. In this evaluation, we find calibration estimates as follows:

$$\hat{t}_k^{CAL} = \sum_{i\epsilon s, i\epsilon k} w_i y_i. \tag{11}$$

where

$$y_i = \sum_{\forall c} y_{ic}. \tag{12}$$

5. Evaluation Design

We used data from the Finance components of the 2002 and 2007 CoG. We restricted our universe to units surveyed in both census years. The 2002 CoG supplied the auxiliary data. The 2007 CoG was the sampling frame in this evaluation. We selected 500 independent samples according to the design detailed in section 2. Each sample contained 5,378 non-certainty units and was used to estimate total revenue by state for non-certainty units in 2007. We compared the MSE for three estimators, Horvitz-Thompson (HT), DB, and calibration for aggregates in each sample.

This evaluation design is similar to the production environment because units that appear in the 2002 CoG but not the 2007 CoG are deaths (governments that no longer exist) and would not be considered in a typical survey year. Units that appear in the 2007 CoG but not in the 2002 CoG are birth units and are treated differently in production for estimation. Focusing only on the units common to both censuses in this evaluation will yield results that can be applied to future survey cycles.

Mean Square Error (MSE)

The MSE provided a composite measure of accuracy and precision. We computed the MSE for the three estimators. Estimates with smaller mean-squared errors were more desirable. From each sample b, we found an estimate of t_k denoted as $\hat{t}_{k_b}$. Then, we estimated the MSE for each of the three estimators as follows:

$$\widehat{MSE}(\hat{t}_k) = \frac{1}{500} \sum_{b=1}^{500} \left(\hat{t}_{k_b} - t_k\right)^2. \tag{13}$$

6. Results

Results from SUDAAN indicated that calibration converged for all iterations and the weight adjustment factor in every sample was centered at one. We estimated 49 state revenue totals for non-certainty units. The estimated MSE of the calibration estimator was the smallest for the majority of the states (26 out of 49). Excluding calibration estimates, the DB estimator performed better than HT.

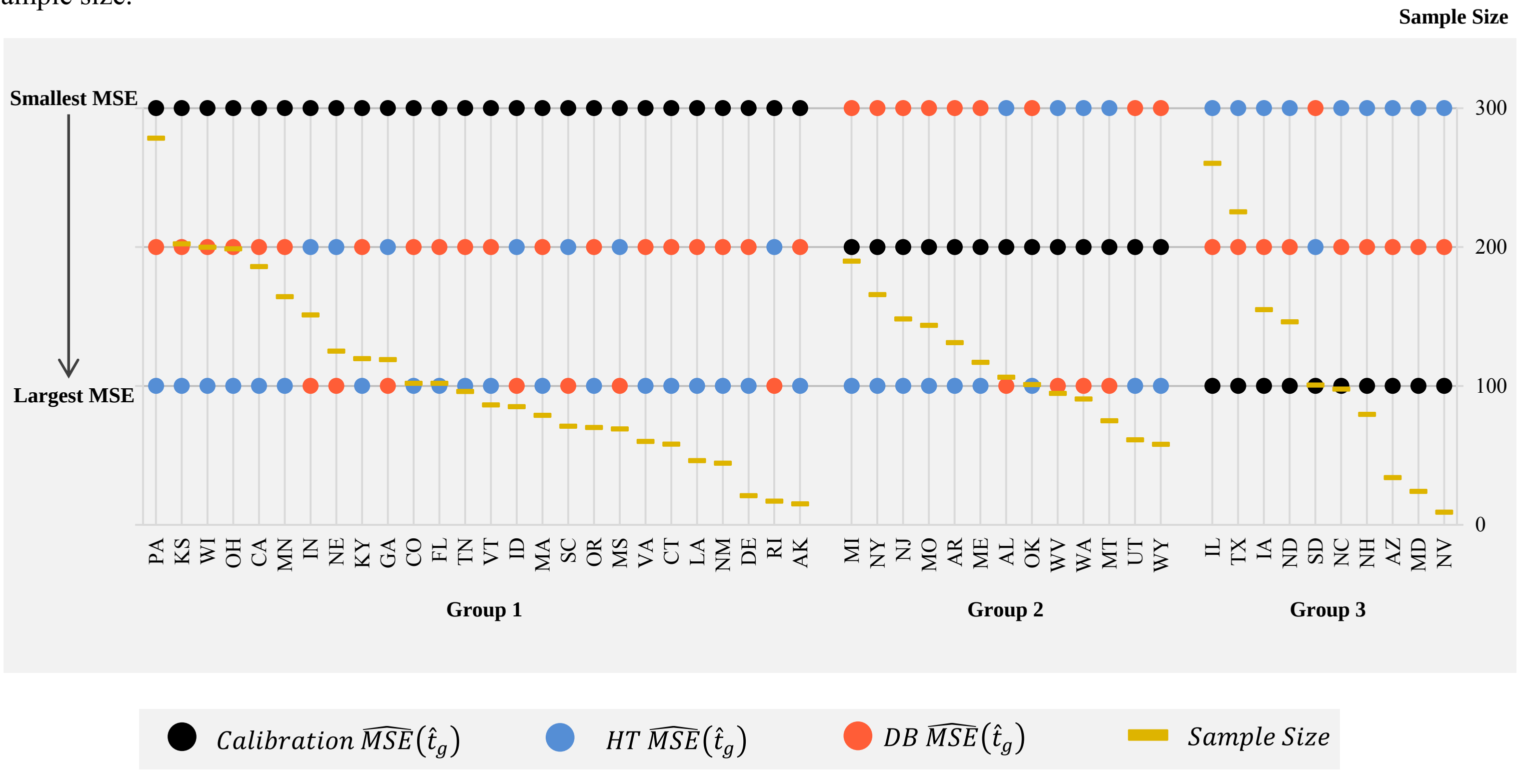
The estimated MSEs from this evaluation spanned a very wide range of values, which made them difficult to chart. Scaling the MSEs using a logarithmic transformation or computing the relative root MSE not only reduced their magnitude but also the magnitude of their differences, which made them harder to distinguish on a graph. To resolve that problem, we plotted their ranks instead. In Figure 1, the top row shows the smallest MSE and the bottom row shows the largest MSE. Figure 1 includes three groups. Group 1 shows the states where calibration had the smallest MSE. Group 2 includes the states where calibration was the second smallest MSE and Group 3 contains the remaining states. Calibration had the smallest estimated MSE for the states in group 1. In group 1, when calibration is the best estimator, the middle row shows that DB outperforms HT. In group 2, calibration is the intermediate performer and DB is the best, having the smallest estimated MSE for most of the states in the second group (9 out of 13). When we compared DB and HT estimates only, DB estimates had smaller MSEs for a majority of states (29 out of 49). When you restrict your attention to states with fewer than 100 units, the calibration estimator performed best when looking at MSEs. Under this restricted focus, calibration has the smallest estimated MSE for 14 out of 24 states.

7. Conclusion

Calibration outperformed HT and DB estimation overall. Even with large sample sizes where the HT estimator is expected to perform very well, calibration estimates tended to have the smallest MSEs. In states with small samples ($n < 100$), calibration performs the best. When comparing HT and DB only, DB estimation is preferred over HT. Future research will include information other than sample size to find patterns in the estimator performance. Additionally, using a vector of totals for the set of calibration equations may improve the MSE of calibration in future studies. Finally, exploring different values for the lower and upper bounds as well as a centering parameter in SUDAAN may also yield calibration estimates with smaller MSEs.

Figure 1. Ranked MSEs and Sample Sizes for Estimates of Total Revenue from Non-Certainty Units by State

Rather than plotting the MSEs, we plotted their rankings. Each of the three rows rank the MSEs from smallest to largest, with the best estimator, i.e. the estimator with the smallest MSE, at the top. The circles represent the estimators and the dashes indicate the sample size.



Source: U.S. Census Bureau, 2002 and 2007 Census of Governments - Finance Component

8. References

Barth, J., Cheng, Y., and Hogue, C. (2009). "Reducing the Public Employment Survey Sample Size," 2009 Joint Statistical Meetings

Cheng, Y., Corcoran, C., Barth, J., Hogue, C. (2009). "An Estimation Procedure for the New Public Employment Survey Design," 2009 Joint Statistical Meetings

Corcoran, C., Cheng, Y. (2010). "Alternative Sample Approach for the Annual Survey of Public Employment Payroll," 2010 Governments Division Report Series, Research Report #2010-5

Deville, J., Särndal, C. (1992). "Calibration Estimators in Survey Sampling," Journal of American Statistical Association

Deville, J., Särndal, C, and Sautory, O. (1993)."Generalized Raking Procedures in Survey Sampling," Journal of American Statistical Association

Kott, P. S. (2011). "WTADJX is Coming: Calibration Weighting in SUDAAN when Unit Nonrespondents Are Not Missing at Random and Other Applications," 2011 Joint Statistical Meetings

Kott, P.S. and Chang, T. (2010). "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse," Journal of the American Statistical Association

Tran, B., Hogue, C. (2009). "Small Area Estimation for Government Surveys," 2012 Joint Statistical Meetings

Research Triangle Institute (2012). SUDAAN Language Manual, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.