Evaluation Study of Calibration Estimation for the Annual Survey of Local Government Finance

Elizabeth L. Love, Bac Tran

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

Abstract

The purpose of this research is to compare the performance of two estimators for the Annual Survey of Local Government Finance (ALFIN): calibration and direct estimator (Horvitz-Thompson). We conducted this evaluation using data from two census years, 2002 and 2007. Samples that replicated the sample design of the current ALFIN were drawn from 2007 census data. The known population totals under different aggregate levels from the 2002 census year were used as calibration totals. The mean squared errors, sampling biases and variances were used as criteria for the evaluation.

Keywords: auxiliary variable, calibration estimation, mean squared error

1. Introduction

The Annual Survey of Local Government Finance (ALFIN) collects data about the revenues, expenditures, debts, and assets of local governments across the United States. Local estimates published from the ALFIN are aggregates of all local government units recognized by the ALFIN: county, municipality, township, school district in conjunction with data collected from the Annual Survey of School System Finances, and special district. Statistics published from the ALFIN are used to estimate the government component of the gross domestic product and provide invaluable information to research organizations, federal agencies, and the public.

We used two different estimation methods for the 2009 and 2010 ALFIN survey cycles. We used decision-based estimation in 2009 to obtain reliable totals. The decision-based approach is based on a decision to combine strata to generate estimates that are more accurate. Results from hypothesis testing the slope from a linear regression model determined whether there was a statistical difference between two strata. When there was no statistical difference between two strata, they were combined, which yielded more accurate estimates. Decision-based estimation yielded precise state estimates of high level aggregates like total revenue but did not generate estimates of low level aggregates with comparable accuracy, see Cheng et al. (2009). In 2010, we explored stepwise-ratios for estimation. We observed that stepwise-ratio estimation performed well for estimates with small yearly changes but did not estimate as accurately data with moderate to extreme changes from year to year. For methodological details, see Tran et al. (2012). While survey analysts researched the feasibility of these two estimation strategies for the ALFIN prior to production, a formal evaluation that concentrated on the ALFIN data had not been conducted.

We considered calibration as an alternative estimation methodology in 2011 as part of an ongoing effort to provide estimates in an efficient timing with considerable quality. The data sources we used for this research were the 2002 and 2007 Censuses of Governments: Finance. We conducted an evaluation to compare the performance of the estimators: Calibration and Horvitz-Thompson (HT).

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

2. Calibration Estimation Methodology

Calibration methods consist of reweighting units (i.e. adjusting the survey weights) so that survey estimates of totals coincide with known population totals from external sources, also called auxiliary information. Calibration estimators rely on auxiliary information to adjust the sampling weights with regard to a set of restraints called calibration equations. This paper follows the notation in Särndal and Deville (1992).

Consider a finite population $U = \{1, \dots, k, \dots, N\}$. Let a probability sample s ($s \subseteq U$) be drawn with a given sampling design. Assume that the inclusion probabilities $\pi_k = \Pr(k \in s)$ and the joint inclusion probabilities $\pi_{lk} = \Pr(k \,\&\, l \in s)$ are always positive. These assumptions become more important when estimating the variance. Let y_k be the value of the variable of interest, y, for the k^{th} population element in U. Let x_k be the value of the auxiliary variable, x, for the k^{th} population element, x_k can be a vector containing many variables but this research focuses on a single variable.

Suppose we observe (y_k, x_k) and assume that the population total, $t_x = \sum_U x_k$, is known. The goal is to estimate the population total t_y by adjusting sample design weights, $d_k = \frac{1}{\pi_k}$, to be w_k,

$$t_x = \sum_s w_k x_k. \quad (1)$$

There are varieties of estimators that satisfied this condition. See Deville and Särndal (1992). Deville, Särndal, Sautory (1993) suggested choosing adjusted weights that meet (1) and close to the survey weight. They called this class of estimators calibration estimators. The closeness of the calibrated weights to the survey weights can be measured by a distance function $G(\frac{w_k}{d_k})$ where G is strictly non-negative and convex. G also has G(1) = G'(1) = 0 and G"(1)=1. The total distance for the whole sample is $\sum_s d_k G(\frac{w_k}{d_k})$. The goal is to minimize $\sum_s d_k G(\frac{w_k}{d_k})$ subject to the constraint (1). Deville and Särndal (1992) and Deville, Särndal, Sautory (1993) discussed a variety of distance functions. In our research we used the linear method, defined as $G\left(\frac{w_k}{d_k}\right) = \frac{1}{2}\left(\frac{w_k}{d_k} - 1\right)^2$. The linear method can yield a set of calibration weights with undesirable properties like weights that fall below one or worse, become negative. Motivated to find a set of calibration weights that do not contain extreme values when adjusting the survey weight, statisticians explored alternative distance functions. See Deville and Särndal (1992). In their derivation of the calibration estimator, Deville and Särndal proved that calibration is equivalent to generalized regression estimation (GREG) when the linear method is used. In the same work, they proposed a variance estimator:

$$\hat{V}\left(\hat{t}_{ygreg}\right) = \sum_{k,l \in s} \sum_{k,l \in s} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}}\right)(w_k e_k)(w_l e_l) \quad (2),$$

where the e_k are residuals from fitting the generalized regression model where the auxiliary variable is the only predictor for y_k. Deville and Särndal provided a detailed derivation of this variance estimator.

3. Study Design

The general sample design used for the ALFIN remained the same from 2009 to 2011 but the total number of units changed because births (newly created governments) were added and deaths (governments that closed) were removed annually. Certainty units accounted for about 55% of the sample. Criteria for certainty units included large-population general-purpose governments (cities, counties, townships) and special districts with sizeable amounts of long-term debt, revenue, and expenditure reported in the 2002 Census of Governments: Finance (COG:F). Less than 1 percent of the units (~103 units) included in the sample were births (newly created governments) or non-activities, governments with (1) no reported debt or expenditure or (2) reported values of zero for both debt and expenditure. Births and non-activities were selected by simple random systematic sampling. We selected the remainder of the units with a probability proportional-to-size (πps) sample design. For strata with a substantial number of small townships or special districts, the stratum was further divided with a cutoff C below

which only a proportion of the units were selected to remain insample, (see Barth, 2009). The size variable was the maximum of total expenditure and long-term debt in 2002.

The samples selected for this research replicated the same selection process. The 2002 COG: F data supplied the auxiliary information needed to select the samples for this simulation study. The sampling frame was the government units surveyed in the 2007 COG: F. We replicated 100 samples and each sample contained 10,875 units to represent four types of local governments: counties, municipalities, townships, and special districts.

3.1. Data

The 2002 and 2007 Censuses provided the data sources for this research. The 2002 Census data supplied the auxiliary information used for this evaluation. The total, t_x in (1), was based on 2002 Census data. We used the 2007 Census data as the sampling frame from which 100 replicated samples were drawn. Estimates from each evaluation sample were compared to known 2007 Census totals.

Data collected in the ALFIN can be broken down into four major categories: revenue, expenditure, assets, and debt. For simplicity, this study included revenue data in the local governments only. The estimates generated in this evaluation were the same as the statistics published from the ALFIN, which were the aggregation of item codes. Each item code represents a different financial activity of a government. Statistics from the ALFIN are estimates of totals based on the aggregation of one or more item codes. The statistics published from the ALFIN contain nested estimates. We excluded the nested estimates and only kept the estimates at the lowest level of aggregation. Other exclusions were aggregates completely comprised of certainty units. We excluded those estimates because they have no sampling bias or sampling variance.

We excluded data collected from the District of Columbia because it does not have any non-certainty local governments in sample. We also removed data from Hawaii because its local government units only, provided data on intergovernmental revenue. All other revenue data in Hawaii come from local certainty units or the state government.

Financial support from other governments, taxes, charges, utilities, and employee-retirement comprised revenue sources for local governments. The Annual Survey of Local Public Pensions provided the data that generated local estimates for employee retirement and unemployment compensation. Since this evaluation only focused on data collected by the ALFIN, those insurance trust data were removed from this evaluation.

Intergovernmental revenues are amounts received from other governments. Local governments received funding from the federal and state governments. This evaluation included data about those intergovernmental revenues because local units reported that information. State and federal governments receiving revenue from local governments were not reported by local units and consequently were excluded from this analysis.

The data source for this research was revenue data collected from local governments surveyed in the 2002 and 2007 CoGs: F.

3.2. Evaluation Design

We classified all units in the universe as either certainties or non-certainties denoted as $C \; and \; NC$ respectively. Estimates for the ALFIN are given by state and can be grouped by revenue source. There was a calibration total for every state s and each revenue source i (see Table 1) which we denoted as t_{si} and is defined as follows for certainty and non-certainty estimates:

$$t_{si} = t_{si}^{C} + t_{si}^{NC}, \qquad where \;\; t_{si}^{C} = \sum_{k \in C_{si}} x_k \;\; and \;\; t_{si}^{NC} = \sum_{k \in NC_{si}} x_k.$$

Revenue data from the 2002 Census supplied the auxiliary information, x_k, needed to estimate t_{si}^{NC}. The table below lists every revenue item code considered in this evaluation as well as its item code group assignment.

Table 1. Revenue Item Code Groups used for Calibration Totals

i	Group Name	Item Codes
1	Intergovernmental revenue from Federal Gov't	B01, B21, B22, B30, B42, B46, B50, B59, B79, B80, B89, B91, B92, B93, B94
2	Intergovernmental Revenue from the State Gov't	C21, C30, C42, C46, C50, C79, C80, C89, C91, C92, C93, C94
3	Taxes	T01, T09, T10, T11, T12, T13, T14, T15, T16, T19, T20, T21, T22, T24, T25, T27, T28, T29, T40, T41, T50, T51, T53, T99
4	Charges	A01, A03, A09, A10, A12, A16, A18, A36, A44, A45, A50, A59, A60, A61, A80, A81, A87, A89
5	Miscellaneous General Revenue	U01, U11, U20, U30, U40, U41, U50, U95, U99
6	Utilities and Liquor Store*	A90, A91, A92, A93, A94
*We grouped utilities revenue and liquor store revenue together because liquor store revenue is mainly a state government revenue.		

3.3. Evaluation Criteria

Estimates from calibration and HT were evaluated on the following criteria:

Mean Square Error

The MSE provided a composite measure of accuracy and precision. We computed the MSE for both estimators. Estimates with smaller mean-squared errors were more desirable.

$$MSE(\hat{t}_{si}) = \widehat{Var}(\hat{t}_{si}) + \left(\widehat{Bias}(\hat{t}_{si})\right)^2 = \widehat{Var}(\hat{t}_{si}) + \left(\hat{t}_{si} - \sum_{k \in U_{si}} x_k\right)^2$$

Relative bias given variance threshold

In practice, survey analysts usually have a maximum variance that should not be exceeded. In our research, we applied a 30% CV threshold to all estimates. For estimates that did not exceed the threshold, we computed the relative sampling bias for both estimators as

$$\widehat{Relative\ Bias}(\hat{t}_{si}) = \frac{\left(\hat{t}_{si} - \sum_{k \in U_{si}} x_k\right)}{\sum_{k \in U_{si}} x_k}.$$

If the CV of both estimators did not exceed 30%, then the estimator with the smaller sampling bias was preferred.

4. Results

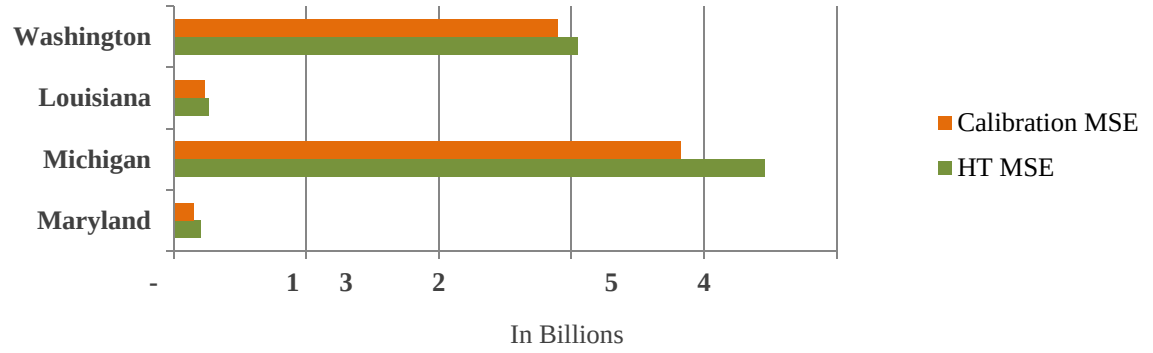
Calibration estimates tended to have lower MSEs for all 100 samples. Given the 30% CV threshold, calibration estimates were more likely to have smaller sampling biases. The charts in this section show some of the results from this evaluation. Rather than showing results from a single sample, the charts show values averaged over 100 samples. We selected the states in each chart because their values were of the same order of magnitude which makes it easier to view them all on the same scale. There were six item code groups. The two figures show results from hospital charges and water supply revenue item code group.

Mean Squared Errors

We compared estimates from calibration and HT for 100 replicated samples. About 70% of calibration estimates from every sample had smaller MSEs than the HT. The squared sampling bias accounted for more than half of the MSE for about two-thirds of the estimates for HT and calibration. Figure 1 shows the MSE of hospital charges for selected states averaged over 100 samples. In the chart below, the MSE is smaller for the calibration estimates than the HT estimates for the four states shown. Even in states like Michigan and Washington where the MSEs were

smaller compared to other states, the calibration estimate continued to outperform the HT. The results in Figure 1 were consistent with outputs from each sample.

Figure 1. Selected MSE Estimates of Hospital Charges Averaged Over 100 Samples
Comparing Mean Square Errors from HT and Calibration Estimates



Source: 2002 & 2007 Censuses of Government: Finance

Relative bias given variance threshold

The threshold for CVs for both estimators was 30%. The total number of estimates per sample was about 950. Exceeding the variance threshold was a little more likely for calibration than it was for HT, but overwhelmingly (at least 95% of all) estimates did not exceed the CV threshold for both estimators in every sample.

Looking only at estimates that did not exceed the variance threshold for both estimators, calibration had smaller relative biases for 80% - 86% of the estimates. Figure 2a compares calibration and HT estimates of water supply revenue to the known 2007 Census total for selected states. Figure 2b shows the relative bias for water supply estimates for the same states that appear in Figure 2a. Together, the charts show that the calibration estimator has smaller bias than the HT estimator. While the difference between an estimate and the actual total is important, adding the relative bias as a percentage provides additional information to help us determine not only which estimator performs better but also gives insight into how much better. The relative sampling bias for calibration estimates was lower. It is important to remember that these dollar amounts are in thousands because smaller differences can represent hundreds of thousands of dollars. For example, Colorado's estimates in Figure 2a showed that the calibration estimate is slightly higher than the HT estimate but the two estimates are more than $20 million apart. Figure 2b illustrated that the HT estimate for Colorado's water supply revenue is almost five percent smaller than the actual total.

Figure 2a. Selected Estimates of Water Supply Revenue Averaged over 100 Samples
Comparing HT and Calibration Estimates to Actual Totals

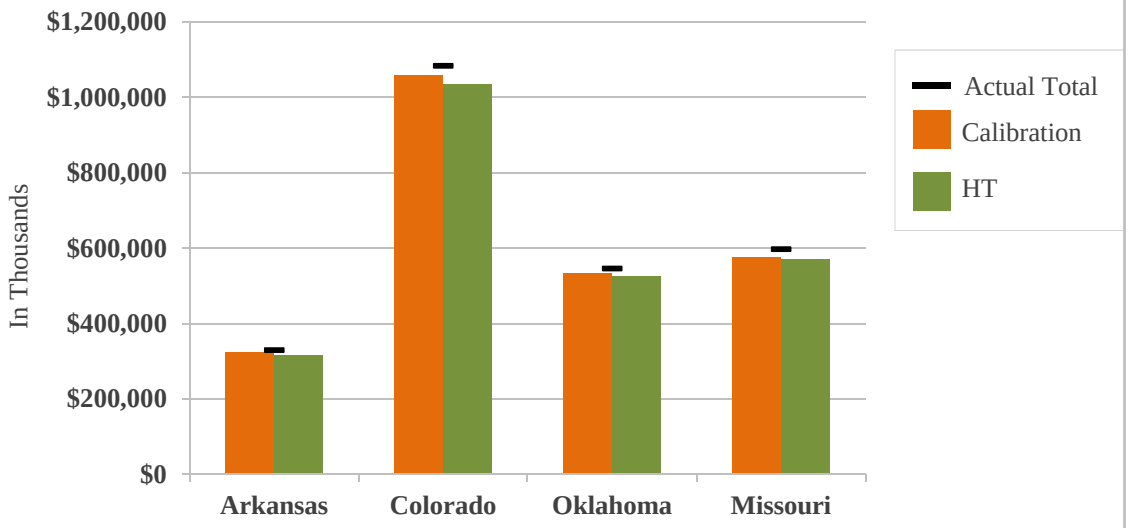
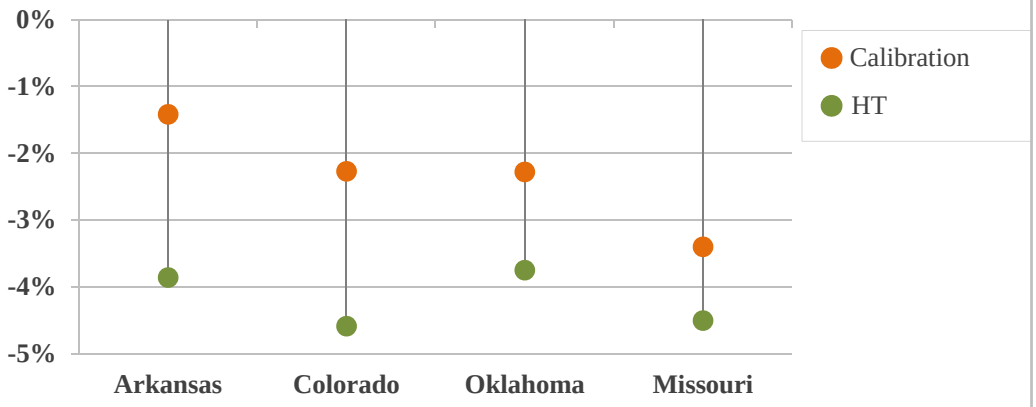


Figure 2b. Relative Sampling Bias for Selected Water Supply Revenue Estimates Averaged over 100 Samples
Comparing the Relative Bias of HT and Calibration Estimates



Source: 2002 & 2007 Censuses of Government: Finance

5. Conclusion

Findings in this evaluation indicated that calibration estimation performed better than HT, as expected. How to quantify *how much better* became the more critical issue. The metrics we used to quantify the differences between the two estimators were MSEs, the relative bias given a variance threshold.

For a majority of the estimates, the MSEs were smaller for calibration. Another important observation was that the squared bias tended to be the larger component of the MSE for the calibration estimates whereas the variance was usually the primary contributor for HT estimates. While the MSE may be the usual indicator that survey analysts use to determine which estimate is better, other metrics may prove to be better in practice given the guidelines at different statistical agencies.

More than 95% of all estimates were under the CV threshold for both estimators. Focusing only on those estimates that did not exceed the CV limit, calibration estimates had smaller biases for at least 80% of all estimates in each

sample. Using the MSE alone, calibration outperformed the HT estimator for 70% of all estimates but applying the Census Bureau standard, calibration yielded better estimates for 76% of all estimates. The metric used is an important determinant when quantifying how much better. In this analysis, calibration was the better estimator.

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

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. 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