Using Administrative Data to Improve Survey Estimation The 1997 Survey of Minority-Owned Business Enterprises

Richard A. Moore, Jr. and Carol V. Caldwell, US Bureau of the Census

Richard A. Moore, US Bureau of the Census, Room 1172-FB3, Washington, DC 20023
Ph: 301-457-3310, Fax: 301-457-3396, e-mail: Richard.A.Moore.Jr@ccmail.census.gov

1. Abstract

The Survey of Minority-Owned Business Enterpises (SMOBE) is a sample survey conducted every five years which provides estimates of the number, receipts, employment, and payroll of minority-owned U.S. companies. The sample for SMOBE is drawn from a universe of businesses stratified by state, industry, and the presumed race of owner(s). Selected businesses are canvassed for the race and ethnicity of the owners. Traditionally, each response is combined with administrative receipts, payroll, and employment information and then weighted by the inverse of the probability of selection to produce estimates by state, industry, and minority group. Design-based variances are produced for each estimate.

While significant improvements to SMOBE have been made since the basic design was adopted in 1982, there are still problems which have prevented us from meeting all of the increasing demands of the data users for more accurate estimates at finer levels of detail. This paper will first briefly discuss the basic design and its problems. It will then focus on a shift from the traditional simple weighted estimator to a ratio estimator for the purpose of providing more accurate detailed estimates.

2. Introduction: The Basic Frame Creation and Traditional Estimation Approaches

General Information. SMOBE is a sample survey conducted once every five years in conjunction with the quinquennial Economic Census. It is the most comprehensive source of basic economic statistics on businesses owned by people of Black, Hispanic, Asian-Pacific Islander, or American Indian-Alaskan Native ancestry. The survey is based on the company rather than its establishments. Published data include the number of firms, gross receipts, number of paid employees, and annual payroll tabulated by geographic area, industry, firm size, and legal form of organization.

Universe Creation. The 1997 SMOBE universe is defined as each company (1) which operates in any industry (except Agricultural Production, Government, the U.S. Postal Service, Household Services, and membership organizations), (2) which is not foreign-owned (i.e., it is based in one of the 50 states or DC), (3) which had non-zero payroll or receipts in excess of $1,000 in 1997, and (4) which was required to file a tax return to the Internal Revenue Service (IRS) for the 1997 tax year. For each such company, the IRS provides the Bureau of the Census with information from which we obtain its geographic location, its principle industry, and its basic economic variables (employment, payroll, receipts). For all sole proprietorships and many of the partnerships and corporations, the IRS also provides the Bureau with the Social Security Number (SSN) of each owner. The Social Security Administration (SSA) then provides the Bureau with the race and surname information supplied to them by the individual on his application for the SSN.

From this information, the Census Bureau constructs a frame of approximately 19 million different companies that are inscope to the 1997 SMOBE. This frame contains all the information necessary to produce the minority-owned business statistics except race and ethnicity information of the ownership. The purpose of SMOBE is to obtain this information for a representative sample of these companies. The sample must be large enough and allocated in a manner that allows us to make accurate minority business estimates for each state by industry (2-digit Standard Industrial Classification, SIC) class.

Race of Ownership Inference. Unfortunately, minority-owned businesses constitute about 12 percent of most of the over 33,000 different state by industry strata. To produce accurate minority businesses statistics at this level would require a sample in excess of 5 million companies. To alleviate this problem, we use 15 different sources of administrative data to identify businesses more likely to be minority-owned . Moore and Williams (1998) describe this procedure in detail. A race inference for each firm was determined on the basis of

1. the firm's race and ethnicity responses to a previous SMOBE
2. the firm's appearance on a publicly disseminated minority business list
3. the owner's self-designated race affiliation on his SSN application
4. the owner's country of birth from the SSN application
5. various surnames supplied by the owner on his SSN application
6. strings in the firm name which indicate a possibility of minority ownership
7. race distributions for various ZIP Codes in which the firm operates
8. race of ownership distributions for various state-industry groups in 1992

Sampling, Estimation, and Variance. All eligible companies are stratified by the race inference, state of operation, and industry classification. A sample is selected from each stratum and canvassed for race and ethnicity of ownership. From the set of respondents, race and ethnicity estimates are derived for each stratum. The estimates and variances are then aggregated over the appropriate strata to obtain publication cell estimates and variances. Suppose, for example, we wanted to produce minority-owned firm count and receipts estimates for Publication Cell D, let

D = Union of Stratum 1, Stratum 2, ..., Stratum H;
X_D = the minority-owned firm count estimate for Cell D;
R_D = the aggregate minority-owned receipts estimate for Cell D;
N_h = the number of firms in Stratum h;
n_h = the number of firms selected from Stratum h;
r_h = the number of responses received from Stratum h;
m_h = the number of firms who responded as minority-owned in Stratum h;
R_{hi} = receipts of the i-th firm in Stratum h;

then

$$X_D = \sum_{h=1}^{H} N_h \times \frac{m_h}{r_h}, \qquad \textbf{(1)}$$

and

$$R_D = \sum_{h=1}^{H} \left(\frac{N_h}{r_h} \cdot \sum_{i=1}^{m_h} R_{hi}\right) = \sum_{h=1}^{H} \left(N_h \cdot \frac{m_h}{r_h} \cdot \sum_{i=1}^{m_h} \frac{R_{hi}}{m_h}\right) \qquad \textbf{(2)}$$

The accompanying designed-based variances are

$$VAR(X_D) = \sum_{i=1}^{H} (N_h)^2 \cdot \left(1 - \frac{r_h}{N_h}\right) \cdot \frac{\frac{m_h}{r_h} \cdot \frac{(r_h - m_h)}{r_h}}{r_h} \qquad \textbf{(3)}$$

and

$$VAR(R_D) = \sum_{i=1}^{H} (N_h)^2 \cdot \left(1 - \frac{r_h}{N_h}\right) \cdot \frac{\sum_{i=1}^{m_h} \frac{R_{hi}^2}{r_h} - \left(\sum_{i=1}^{m_h} \frac{R_{hi}}{r_h}\right)^2}{r_h}. \qquad \textbf{(4)}$$

3. Estimation Problems with This Design

Although this design has been used since SMOBE migrated to a domain-based estimation methodology with the 1982 survey, it has several major shortcomings which limit the accuracy of the published estimates.

Problem #1. Coverage. Some minority-owned firms are virtually impossible to pre-identify. Traditionally, SMOBEs have sampled this group of businesses at rates lower than 1 in 100. Although this sample provides an estimate for the number of minorities which were not pre-identified, it is not large enough to support detailed estimates.

Problem #2. Scarcity of Minority-Owned Businesses. Minority-owned businesses are extremely rare. Although the 1997 SMOBE mailed questionnaires to about 2.5 million firms and received approximately 2.0 million completed questionnaires, only a small percentage (about 12 percent) of these respond as minority-owned. To provide accurate firm counts at the stratum level, most strata have to be sampled at high rates. Of the 33,384 noncertainty strata, the median sampling rate is 1 in 3.25. About 95 percent of the strata are sampled at a rate lower than 1 in 40, with the lowest rate being 1 in 500.

Problem #3. Inaccuracy of the Race Classification. It is difficult to accurately stratify each business. Although the employer and nonemployer registers, which are created for the Economic Census, contain accurate state and industry information, very little accurate data are available on the races of the owners of each firm. Sample sizes to achieve target variances are determined under the assumption that firms are stratified correctly. When they are not, the current methodology will inflate the design-based variance. The magnitude of the inflation is related to the weights of the incorrectly classified cases. One or two of incorrectly classified cases, with large weights, can cause a substantial increase.

Problem #4. Non-response Bias. The response rates for minority-owned businesses are presumed to be lower than those for non-minority-owned firms. From Equation (1), one sees that such a problem underestimates the estimated proportion of minority-owned firms in each stratum ($p_h = m_h / r_h$). This results in a negative bias in the firm count estimates at all levels of aggregation.

Problem #5. Estimation of Auxiliary Statistics (Aggregate Receipts, Employment, Payroll). The sampling rates for the individual strata are determined by variance constraints placed on the firm count estimates for various domains of interest. The estimator for the auxiliary variables in Equation (2) usually produces estimates with significantly higher variation. (To provide some control, cases with extremely large receipts are selected with certainty.) For the 1992 SMOBE, the relative standard errors (RSEs) on the state-level Hispanic-owned firm count estimates ranged from 0 to 13 percent, while the RSEs on the corresponding aggregate receipts estimates ranged from 2 to 29 percent. Most auxiliary estimates at the sub-state level have RSEs so large that they are virtually meaningless to the data users who insist on accurate estimates.

4. Goals of the 1997 SMOBE Redesign

Having identified these problems with the 1992 SMOBE methodology, the Census Bureau developed an approach for the 1997 survey which addressed these problems. This approach had the following five main goals:

- eliminate as much of the coverage loss as possible,
- provide accurate firm counts estimates at as fine a detail as possible,
- develop an estimator which accounts for inaccuracies of the race classification,
- use statistically sound adjustment procedures to reduce non-response bias, and
- reduce the RSEs of the auxiliary variable estimates.

5. The 1997 SMOBE Redesign Strategy

To achieve these goals, the 1997 SMOBE redesign employed three new steps:

- expanding the sources of administrative data to pre-identify more firms as possibly minority-owned,
- using historical data to quantify the likelihood that each inference is correct, and
- incorporating that likelihood in the sample design and estimation process.

This section describes the contributions of each step.

Expanded Sources of Administrative Data. Previous SMOBEs obtained their race inferences from (1) race and ethnicity responses to a previous SMOBE, (2) the self-designated race on the SSN application, and (3) the various surnames on the SSN application. For 1997, we added inferences from (4) publicly disseminated lists, (5) the owner's country of birth, (6) strings in the firm name, (7) race distributions for various ZIP Codes in which the firm operates, and (8) 1992 race of ownership distributions for various state-industry classes.

For the 1997 survey, the original three sources assigned race inferences to 6.62 million firms. Approximately 1.76 million of these are minority-owned. From the 12.36 million firms for which no inference could be made, approximately 0.48 million minority-owned firms were missed. The five new sources of data provided inferences for an additional 4.95 million of these, of which about 0.35 million are minority-owned. The additional sources reduced the undercoverage from 0.48 to 0.13 million minority-owned firms.

Addition of a Likelihood Estimator Field. The 1997 SMOBE frame construction made race inferences from a variety of sources. Some of the sources are extremely good predictors of the race of the owner (e.g., the firm's response to the 1992 SMOBE), while others provide only minor evidence of minority-ownership (e.g., inferences based on the distribution of minority-owned businesses in the 1992 SMOBE). Each sampling record was not only assigned a race inference, but also a likelihood that the inference was correct. The likelihood enables us to quickly identify the most likely minority-owned operations. Table 1 below shows this likelihood distribution in intervals of 0.200.

Table 1. Distribution of the Likelihood That the Race Inference Is Correct
In Intervals of 0.200
Percentage of Sampling Units

Likelihood Range	Overall	Sole Props	Non Sole Props	Phase I Universe	Phase II Universe
0.000 - 0.199	84.4	82.1	89.9	91.7	82.2
0.200 - 0.399	2.2	0.8	5.4	4.9	1.4
0.400 - 0.599	2.4	2.5	2.0	2.0	2.5
0.600 - 0.799	1.7	2.1	0.6	0.6	2.0
0.800 - 1.000	9.4	12.5	2.0	0.9	11.9

Table 1 also provides the likelihood distribution for sole proprietorships and non-sole proprietorships. Notice that the distribution for sole proprietorships is more disperse, since we can more easily and accurately assign race of ownership to single owner firms than to most multiple owner (partnerships and corporations) firms.

The 1997 SMOBE canvassed over 2.5 million units. Because of the large size, the universe was divided into two mail-outs, Phase I and Phase II. The Phase I universe consisted of 3.9 million inscope corporations and partnerships that were in business on Dec. 31, 1996. Sole proprietorships were not included, because these have a much higher fatality rate than corporations and partnerships. The Phase II Universe consisted of all 15.0 million inscope firms not included in the Phase I Universe. Although race information was collected from some Phase I Universe firms which ceased operations at the end of 1996, only firms on the 1997 registers were eligible for tabulation in the 1997 SMOBE. Table 1 includes the likelihood distribution for the Phase I Universe, because it was used as a pilot to test the proposed estimation procedure.

Use of the Likelihood in the Estimation Procedure. The value of each sampling unit's likelihood is estimated by making the same inference on the set of 1992 SMOBE respondents and determining the percentage of cases where the race of ownership inference is correct. If conditions have not significantly changed over the past 5 years, and if the set of respondents was representative of the universe in 1992, then we can assume that the likelihood is, at worst, a slightly biased estimate for the probability that the firm will respond as minority-owned (i.e, if $d_{hi} = 1$, if the firm is minority-owned and $d_{hi} = 0$ otherwise, then $\succ(d_{hi}) = l_{hi} + b_{hi}$, with each bias, b_{hi} . 0. Here $\succ(d_{hi})$ denotes the average of d_{hi} of all firms that have the same likelihood). Suppose

D = Union of Stratum 1, Stratum 2, ..., Stratum H;
X_i = the minority-owned firm count estimate for D with Estimator i;
N_h = the number of firms in Stratum h;
n_h = the number of firms selected from Stratum h;
r_h = the number of responses received from Stratum h;
m_h = the number of firms who responded as minority-owned in Stratum h;
l_{hi} = estimated likelihood that the i-th firm in Stratum h is minority-owned; and
d_{hi} = 1, if the firm is minority-owned (0, otherwise).

We can produce several estimators for the number of minorities in Domain D.

Estimator #1. Enumerate the entire universe.

$$X_1 \; ' \; \mathop{\mathbf{j}}_{h'1}^{H} \mathop{\mathbf{j}}_{i'1}^{N_h} d_{hi} \qquad \textbf{(5)}$$

Estimator #2. Tabulate the estimate from the universe by summing the likelihoods.

$$X_2 \; . \; \mathop{j}_{h'1}^{H} \mathop{j}_{i'1}^{N_h} l_{hi} \; ' \; \mathop{j}_{h'1}^{H} \mathop{j}_{i'1}^{N_h} ?(d_{hi}) \qquad \textbf{(6)}$$

Estimator #3. (Estimator of Equation (1)). Select a stratified sample. Consider the set of respondents in Stratum h representative of all cases in the stratum.

$$X_3 = \sum_{h=1}^{H} \sum_{i=1}^{r_h} \frac{N_h}{r_h} \times d_{hi} \qquad \textbf{(7)}$$

Estimator #4. Select a stratified sample. Consider the set of respondents in Stratum h representative of all cases in the stratum. Sum the likelihoods of each selected case.

$$X_4 \approx \sum_{h=1}^{H} \sum_{i=1}^{r_h} \frac{N_h}{r_h} \times l_{hi} = \sum_{h=1}^{H} \sum_{i=1}^{r_h} \frac{N_h}{r_h} \times ?(d_{hi}) \qquad \textbf{(8)}$$

Let’s examine some properties and relationships of each these estimators to each other.

- X_1 is obviously an <u>unbiased</u> estimate for the minority-owned firm count in Domain D.
- X_2 is an implicitly biased estimate of X_1 . The magnitude of its bias is a function of the magnitudes of the bias that each likelihood can predict the corresponding response.
- X_3 is also an implicitly biased estimate of X_1. The magnitude of its bias is a function of the magnitudes of the response bias at the stratum level.
- X_4 has two implicit biases. It combines the implicit biases of the likelihood estimate with those of the response estimate.
- X_2 / X_1 measures the net bias of the likelihood estimation on X_2.
- X_4 / X_3 estimates the net bias of the likelihood estimation on X_4.

Conjecture: $X_4 / X_3 \approx X_2 / X_1$, which implies $X_1 \approx X_2 * X_3 / X_4$, which motivates the following estimator.

Estimator #5. Tabulate the estimate from the universe by summing the likelihoods. Use Estimators X_3 and X_4 to adjust for the likelihood bias of the estimate.

$$X_5 = \sum_{h=1}^{H} \sum_{i=1}^{N_h} l_{hi} \times \frac{\displaystyle\sum_{h=1}^{H} \sum_{i=1}^{r_h} \frac{N_h}{r_h} \times d_{hi}}{\displaystyle\sum_{h=1}^{H} \sum_{i=1}^{r_h} \frac{N_h}{r_h} \times l_{hi}} \qquad \textbf{(9)}$$

Note: Suppose RCT_{hi} were the receipts of the Firm i in Stratum h. The companion receipts estimate would be

$$X_5^{*} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} l_{hi} * \frac{\sum_{h=1}^{H} \sum_{i=1}^{r_h} \frac{N_h}{r_h} * d_{hi} * RCT_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{r_h} \frac{N_h}{r_h} * l_{hi} * RCT_{hi}} \qquad \textbf{(10)}$$

6. Potential Advantages of the Ratio Estimator

The ratio estimators, X_5 and X_5^*, are the tools which provide the initial steps for attaining many of the estimation goals for SMOBE in 1997 and beyond. From both a theoretical and a logical perspective, they exhibit promise that we can achieve more accurate estimates than the traditional estimator, X_3. These are explained in more detail below.

1. *Compensation for Non-response Bias.* Provided the net bias associated with the estimated likelihoods is independent of the bias associated with the inability of the response set to accurately represent the initial sample, we have constructed an estimator which has compensated for both the non-response and the likelihood biases.
2. *Compensation for Inaccuracies in the Race of Ownership Classification.* Each eligible unit is assigned a small positive likelihood of having an incorrect race inference (l_{hi} . 0). Since we are not limiting our estimates to one stratum per cell ($H \geq 1$), we have constructed an estimator which compensates for inaccuracy in the race of ownership assignment procedure.
3. *More accurate estimates at the publication level.* When H = 1 and

   ?(d_{hi}) = l_{hi},
   S_L^2 = VAR({l_{hi}}), and
   S_D^2 = VAR({d_{hi}}),

   then

   $$?[VAR(X_5)] = ?[VAR(X_3)] - N_h^2 * (1 - \frac{r_h}{N_h}) * \frac{S_L^2}{r_h} * (1 - O(\frac{1}{N_h})) . \qquad \textbf{(11)}$$

   If all assumptions are met as N_h becomes large,

   [VAR(X_5)] . (1 - S_L^2/ S_D^2) * VAR(X_3),

   thus the gain in precision increases as S_L approaches S_D. We hope that these results will generalize to the collapsed strata situation.
4. *A<u>ccurate</u> estimates at finer detail.* Although Equations (9) and (10) use the design stratum (race inference by state of operation by 2-digit Standard Industry Classification (SIC) code) as the basic estimation unit, this method may also be used for estimation at finer levels (e.g., estimation at sub-state levels or at other levels of industrial classification). There is no reason to suspect that the result of Equation (11) won't also apply at any sub-stratum level.

5. *Accurate estimates for the auxiliary statistics.* The result of Equation (11) holds for estimating auxiliary statistics receipts, payroll, and employment. For example, to estimate receipts replace $\{d_{hi}\}$ with $\{d_{hi}*RCT_{hi}\}$ and $\{l_{hi}\}$ with $\{l_{hi}*RCT_{hi}\}$ in the formulas for X_3 , X_4, and S_L.

7. Testing of the Ratio Estimator

As stated in Section 5, SMOBE had two mail-outs. Responses to the Phase I sample were used to provide estimates for various sub-domains of the Phase I Universe. This universe consisted of all corporations and partnerships that conducted business operations on Dec. 31, 1996. A probabilistic sample of approximately 1.0 million corporations and partnerships were selected from this universe and mailed a 1997 SMOBE questionnaire. Over 83.7 percent of these provided race and ethnicity responses for their ownership.

Using both the traditional, X_3, and the ratio, X_5, estimators, we estimated minority-owned firm count and receipts for 3,617 sub-domains which had positive variability. Table 2 compares the two estimators. It shows that, on average, the ratio estimator increases the firm count and decreases the receipts estimates by 1.6 and 3.7 percent, respectively. This provides some circumstantial evidence that the standard estimator does underestimate the true number of minority-owned firms and underestimate their aggregate receipts.

Table 2 also illustrates that the ratio estimator gives a higher relative standard error (RSE) on the firm count estimate for only 1,428 (or 37 percent) of the cells. (RSE estimates for the ratio estimator were calculated using jackknife replication. (Wolter, 1985)) The RSE distribution of the receipts estimates is even more impressive. The ratio estimator gave a larger RSE for only 442 (or 12 percent) of the 3,617 cells. This lends some credibility to the hypothesis that the ratio estimator should reduce the variation for many of the published cells.

Table 2. Comparison of the Standard (X_3) and the Ratio Estimator (X_5) On 3,617 Sub-Domains with Variability

	X_5/X_3		**# Sub-domains**		
Estimate	**Mean**	**Median**	**$X_5 < X_3$**	**$X_5 = X_3$**	**$X_5 > X_3$**
Firm Count	1.016	1.000	1,058	1,460	1,099
Receipts	0.963	0.964	2,360	806	451
RSE	**Mean**	**Median**	**$RSE_5 < RSE_3$**	**$RSE_5 = RSE_3$**	**$RSE_5 > RSE_3$**
Firm Count	1.050	0.985	2,230	59	1,428
Receipts	0.960	0.962	3,100	75	442

We hope to obtain even more impressive results when the ratio estimator is applied to the complete (Phases I and II) set of respondents for several reasons.

- About 80 percent of all minority-owned operations are sole proprietorships. The results from Phase II will dominate the overall results.
- Sole proprietorships generally have lower response rates than corporations and partnerships. Although 83.7 percent of the sampled firms in Phase I responded, we expect response rates of about 75 percent for Phase II. The difference in response rates could result in significantly more non-response bias.
- Under ideal conditions, $VAR(X_5)$. $(1 - S_L^2/ S_D^2) * VAR(X_3)$. Table 2 shows more variability in the distribution of the likelihood of the Phase II Universe than Phase I. For Phase I, S_L^2/ S_D^2 . 0.28; for Phase II, we expect S_L^2/ S_D^2 . 0.73.
- The Phase I Universe contained only corporations and partnerships. We have more extensive administrative information on sole proprietorships. This makes the prediction of race of ownership more accurate for Phase II where almost 90 percent of the universe are sole proprietorships. For many of the same sub-domains, we would expect X_3/X_4 to be closer to 1.000 in Phase II than in Phase I.

8. Conclusion

The paucity of minority-owned firms makes it extremely difficult to make accurate estimates for sub-national and sub-industry domains. With traditional methodology, the universe must be stratified correctly and most strata sampled at extremely high rates. By increasing our ability to process, to make correct race of ownership inferences, to evaluate the accuracy of each inference, and to incorporate this into the estimation procedure, we may be better able to produce accurate estimates for small sub-domains with a reasonable size sample.

9. References

- Cochran, W. G. (1977). Sampling Techniques. John Wiley and Sons, New York, NY.
- Moore, R., Caldwell, C., Detlefsen R. (1998). “Changing the Sample Design to Meet User Needs — The Survey of Minority-Owned Business Enterprises — Past, Present, and Future”, Proceedings of the Section on Survey Research Methodology. American Statistical Society, pages 493-498.
- Moore, R. and Williams, A. (1998). “Using Administrative Data to Enhance the Sampling Frame for the 1997 Survey of Minority-Owned Business Enterprises”, Proceedings of the Section on Survey Research Methodology. American Statistical Society , pages 499-504.
- Sarndal, C. E., Swenson, B., and Wretman, J. (1992). Model-Assisted Survey Sampling. Springer-verlag, New york, NY.
- Wolter, K. (1985). Introduction to Variance Estimation. Springer-Verlag, New York, NY.

Disclaimer. *This paper reports the results of research undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.*