CHANGING THE SAMPLE DESIGN TO MEET USER NEEDS THE SURVEY OF MINORITY-OWNED BUSINESS ENTERPRISES PAST, PRESENT, AND FUTURE

Richard A. Moore, Carol V. Caldwell, and Ruth E. Detlefsen, United States Bureau of the Census
Richard Moore, U.S. Bureau of the Census, Washington, D.C. 20233

Key Words: Minority-Owned Business, Stratification, Multiple Constraints

Introduction

The Survey of Minority-Owned Business Enterprises (SMOBE) is the only ongoing source of comprehensive statistics that count and describe U.S. businesses by the race and ethnicity of business owners. The program combines survey data on race and ethnicity with administrative data to estimate firm count, employment, payroll, and receipts for minority-owned businesses by industry and geography.

The challenge for the SMOBE program is that businesses owned by minorities constitute a small portion of the business universe and are difficult to identify. A large sample is therefore needed to produce accurate estimates. For prior surveys, a fairly simple sample design was used to produce accurate estimates for levels of relatively general detail. Users now require the 1997 SMOBE to provide estimates at much finer levels of detail. A more complex and efficient sample design was necessary to meet the user needs. This paper describes the development of the 1997 sample design, compares it to the previous design, and addresses some of the possible improvements for the 2002 SMOBE.

SMOBE Methodology, 1969 - 1992

The first SMOBE was conducted in 1969. Surveys have been conducted every five years since 1972 as part of the Economic Censuses.

The Census Bureau has always relied on administrative information from other agencies -- especially, the Internal Revenue Service (IRS) and the Social Security Administration (SSA) -- to pre-identify as many minority owners as possible. Once businesses have been segmented into mutually exclusive race/ethnicity classes, stratified systematic simple random sampling is used to select cases from which the estimates are made. The evolution of the frame assignment and sampling methodology from 1969 to 1992 is described in more detail in Sands (1993). The predominant features of the SMOBE design used in the 1982, 1987, and 1992 surveys are described below.

Frame construction.

The Census Bureau obtained a list of Social Security Numbers (SSNs) for each owner of every establishment that filed either an IRS Form 1040 Schedule C (Sole Proprietorship), Form 1065 (Partnership), or Form 1120S (Special Corporation) tax return. The Census Bureau then obtained race information for each of these owners from the SSA. The Census Bureau first divided all business enterprises into categories based on the SSA race classification of the majority of owners. It then further identified businesses with the majority of owners with surnames on compiled lists of surnames associated with persons of Hispanic, Asian, American Indian, or Alaskan Native descent. Based on this information, businesses were assigned to one of four mutually exclusive but exhaustive frames: possible Black-owned (BLACK), Hispanic-owned (HISPANIC), and other minority-owned (OTHER) businesses, and a NATIONAL frame consisting of all other businesses.

Stratification, Sampling, and Estimation.

Businesses in each frame were stratified by legal form of organization (LFO), employer status (employer vs. nonemployer), frame, state, and 2-digit Standard Industrial Classification (SIC) industry.

All sole proprietorships which responded in the previous SMOBE were identified and selected with certainty but were not mailed report forms. The race and ethnicity for the owners of these businesses were tabulated based on their responses to the previous survey. Their receipts, employment and payroll were tabulated from administrative data for the survey year.

Sole proprietorships in the BLACK frame with a race of Black assigned by SSA were selected with certainty and tabulated as Black-owned. The SSA race classifications have consistently included a Black category, but categories for other minorities were expanded in 1981. We could therefore use the SSA classification to reliably

identify a large portion of Black single-owner businesses, but not sole proprietors in the other minority groups.

The remaining businesses in the four frames were sampled and selected units were sent a common questionnaire, which collected information on both race and Hispanic origin. About 1.1 million businesses were selected for mailing, as follows:

- Businesses with large receipts were selected with certainty. This selection provided some control on the variances of the estimates of the auxiliary characteristics (receipts, employment, and payroll).

- Large samples were drawn from the remaining cases in the HISPANIC and OTHER frames. Stratum sampling rates varied, depending on the expected number of Hispanic-, Asian-, or American Indian-owned businesses in each state or industry. The larger the expected number of minority-owned businesses, the lower the sampling rate. A 10 percent coefficient of variation (CV) for each state and each two-digit SIC was imposed on the estimates from each frame.

- A sample of partnerships and special corporations was drawn from the BLACK frame.

- Relatively small samples were selected from the BLACK sole proprietorships and from the businesses in the NATIONAL frame. Responses were used to estimate the number of Hispanic-owned and other-non-Black-minority-owned businesses present in these frames.

Domain estimation was used to eliminate the bias incurred by inadvertently assigning a business to an incorrect frame. In addition, estimates for White-Hispanic, Asian-Hispanic, American Indian-Hispanic, and Black-Hispanic-owned businesses were computed. Although the Bureau did not publish Hispanic statistics by race, these estimates were aggregated and the published numbers then reflected all Hispanic-owned businesses.

The Need for a Redesign in 1997

Refinements in the SMOBE frame construction and sample design had been made throughout the life of the survey. The design features described above for the 1982 through 1992 surveys eliminated many of the biases inherent in the earlier surveys. However, there were still major components of the design that demanded further attention. First, the survey did not cover "full" corporations (i.e., businesses that file IRS Form 1120 tax returns). In 1992, full corporations accounted for 11 percent of the total number of U.S. firms, and 76 percent of total business receipts.

Second, in past surveys, estimates of minority-owned businesses from the NATIONAL frame were not combined with estimates from the BLACK, HISPANIC and OTHER frames. While measures had been taken to improve the identification and placement of "hard-to-identify" minority-owned businesses into the other frames, the estimated number of minority-owned businesses remaining in the NATIONAL frame was still very large, and had large variability. The poor quality of these estimates prevented their inclusion in the main publication tables. Instead, separate summary tables were produced, showing the estimates of these "missed" minority-owned businesses at the U.S. level only.

In designing the 1997 survey, we aimed to address these coverage and estimation problems. We also set goals to meet user needs for (1) separate estimates of Asian and Pacific Islander-owned businesses and Native American and Alaskan Native-owned businesses, and (2) accurate estimates for finer levels of industry and geography.

Evolution of the 1997 SMOBE Redesign

We took a step-by-step approach to improving the efficiency of the SMOBE design, while striving to meet the additional demands of the data users and sponsors. The steps in the process are described below. The improved design will be accomplished with a mailout to about 2.5 million businesses, supplemented by administrative race and ethnicity information available from an additional 127,000 businesses.

Step 1: Introduce Design Efficiencies. To achieve all the required objectives of the 1997 survey, a sample size of over 7.0 million (or 35 percent of the universe) would have been required using the 1992 methodology. We first identified fundamental changes that could be made to the existing sample design to attain greater efficiency:

1) Eliminate additional restrictions on the sample size. In 1992, no sampling rate was allowed to drop below 1 in 10. For 1997, we dropped this restriction.

2) Eliminate low priority strata. The survey sponsors are primarily interested in estimates for the total number of minority-owned firms for certain geographic areas and industrial classifications. Accurate breakdowns for

employers or legal forms of organization are distant secondary priorities. In the 1992 survey, the 5-level stratification produced many strata with only a few observations. This resulted in sampling rates which were very high. For 1997, we eliminated stratification by LFO and employment status, and stratified only by frame by state by 2-digit SIC. To increase the accuracy of the estimates by LFO and employment status, sampling units within each stratum were sorted by those keys and then systematically sampled.

The 1992 sample size was 1.1 million. We did not measure the direct impact of the above design efficiencies on the 1992 sample size. However, we did determine that the required sample size for meeting all 1997 survey goals would have been 7.0 million instead of 5.9 million had we not implemented these efficiencies.

Once we established these design changes, we used the 1992 universe and survey results to measure the incremental impact of each additional requirement for the 1997 survey. Table 1 shows the effect of each step on the estimated sample size.

Step 2: Produce Estimates Accurate at the State by 2-digit SIC Level. To meet the need for accurate estimates at finer levels of detail, we imposed a 10 percent coefficient of variation at the state by 2-digit SIC-level for the BLACK, HISPANIC, and OTHER frames used in the 1992 design. This increased the sample size from 1.1 million to about 2.3 million.

Step 3: Produce Separate Tabulations for Asian- and American Indian-Owned Businesses. Next, we calculated the increase required for separate Native American-owned and Asian/Pacific Islander-owned business tabulations. This essentially involved breaking the OTHER frame into two separate frames --- ASIAN and AMERICAN INDIAN. Such a change resulted in an additional increase of about 550,000 required sampling units. The sample size now stood at 2.9 million.

Step 4: Add "Full" Corporations to the Universe. In 1992, it was estimated that adding the full corporations to the universe would increase the sample by over 2.0 million. With the design efficiencies that we made in Step 1, the sample size increased by only about 300,000 units. The total required sample size was now just under 3.2 million.

Step 5: Sample Firms with Large Receipts with Certainty. As with previous SMOBEs, sample sizes were determined to meet variance constraints only for the firm count estimates. The auxiliary variables (receipts, payroll, and employment) generally had higher variability. In order to provide some control over auxiliary variables, a method of Glasser (1962) was used to identify firms with abnormally large receipts. These

TABLE 1

1997 SMOBE Sample Size Evolution

		TOTAL
	1992 sample size	1,117,452
STEP 1	Implement design efficiencies	not measured
STEP 2	Impose state by industry CV constraints on BLACK, HISPANIC, and OTHER frames	2,301,669
STEP 3	Separate ASIAN and AMERICAN INDIAN frames	2,850,739
STEP 4	Add full corporations to universe	3,170,731
STEP 5	Sample large receipts cases with certainty	3,172,374
STEP 6	Place CV constraints on BLACK and NATIONAL frames	5,886,706
STEP 8	Relax CV constraints for small estimates: -- ASIAN, AMERICAN INDIAN, and HISPANIC frames	4,458,619
	-- BLACK and NATIONAL frames	3,578,407
STEP 9	Relax CV constraints on BLACK and NATIONAL frames - revert to constraints for frame-state and frame-industry	2,071,214
STEP 10	Add 40 4-digit SICs	2,349,732
STEP 11	Use Chromy_Gen software	2,293,640
STEP 11A	Use Chromy_Gen with added state and industry constraints	**2,330,107**

large receipts firms were selected into the sample with certainty. This increased the sample size only minimally. The total sample size remained under 3.2 million.

Step 6: Impose CV Constraints on the BLACK and NATIONAL Frames to Estimate Frame Misclassification. In 1992, about 90,000 of 11.5 million firms were selected from the BLACK and NATIONAL frames to estimate the number of Asian-, Hispanic-, or American Indian-owned businesses not assigned to the HISPANIC or OTHER frame. The responses indicated that we had misclassified about 38,000 Asian-owned, 91,000 Hispanic-owned, and 61,000 American Indian-owned businesses. Because of the large sampling weights, the firm count estimates and accompanying estimates of the auxiliary characteristics from this component of the sample had huge variances. Consequently, they were not included in the main publication tables. Our research estimated that the NATIONAL and BLACK samples would have to be increased to 2.8 million units to produce estimates with the same reliability as the estimates from the other frames, and incorporate them into all the main publication tables. The total sample size with this requirement amounted to 5.9 million cases.

Step 7: Reevaluate the Design. Funding for the 1997 SMOBE was approved for a sample of 2.5 million businesses. Further work on the design was needed to reduce the sample size from 5.9 million to 2.5 million.

We evaluated our top-priority objectives and decided that it was most important to include the full (Form 1120) corporations in the SMOBE universe, provided that we could find some compromises that would still allow us to produce estimates for small domains as well as separate estimates of American Indian-owned and Asian and Pacific Islander-owned businesses.

Step 8: Relax Variability Constraints for Cells with Small Estimates. A driving component of sample size determination was the projected number of minority-owned businesses in each stratum. Some strata had very few. In these, a large proportion of the units must be selected in order meet a tight reliability requirement. A decision was made to relax the variability constraints for the strata expected to contain few minority businesses. This was accomplished by setting the target standard error for each stratum to be the maximum of 10.0 and 10 percent of the stratum estimate. Thus, strata with less than 100 minority-owned businesses would be less reliable than those strata with more than 100 minority-owned firms.

Using these relaxed constraints, we lowered the sample sizes for the ASIAN, AMERICAN INDIAN, and HISPANIC frames by about 1.4 million, and the sample sizes for the BLACK and NATIONAL frames by about 0.9 million. The total sample size was reduced from 5.9 million to 3.6 million businesses.

Step 9: Relax CV Constraints on Estimates of Minority-Owned Businesses from the BLACK and NATIONAL Frames. To further reduce the total sample size, we eliminated the CV constraints for the frame-state-industry level estimates from the BLACK and NATIONAL frames. We replaced these with 10 percent CVs on each state and each industry within each frame. In other words, the design was changed to produce accurate estimates by state and by industry, but not necessarily by industry within state. This change drastically reduced the sample sizes for the BLACK and NATIONAL frames, from 1.9 million to 0.4 million, while maintaining relatively large sampling rates of about 1 in 20, compared to the rate in 1992, which was less than 1 in 100. The total sample now stood at 2.1 million cases.

Step 10: Add Forty 4-digit SIC Codes. Although we redesigned SMOBE to produce estimates accurate at the frame by state by 2-digit industry (SIC) for the ASIAN, AMERICAN INDIAN, and HISPANIC frames, users often want accurate estimates at the frame by state by 4-digit industry level. The Bureau has never before designed SMOBE to accurately estimate at this level of detail. We used some of the remaining available sampling units to include samples for forty of the more prominent 4-digit SICs.

The addition of this detail complicated the design, adding a new level of stratification and several nested layers of variance constraints to ensure that all variability requirements would be met simultaneously. The addition of the specified 4-digit industries increased the total sample size to just over 2.3 million.

Step 11. Use Multi-constraint Optimal Allocation Software. In 1995, Laura Zayatz and Richard Sigman of the U.S. Bureau of the Census developed software to find minimum-cost samples that satisfy multiple variance constraints. The software, called Chromy_Gen, is based on an algorithm described by Jim Chromy (1987). We used this software to achieve a minimum sample for the nested layers of variance constraints introduced by the addition of the forty 4-digit SICs to the sample design. The total sample size was reduced by about 56,000 units (or 2.4 percent) to just under 2.3 million.

At this time, we imposed 10% CV constraints on each minority-state and each minority-2-digit SIC estimate. These were the only constraints imposed in the 1992 survey. The total number of constraints for 1997 was increased from 24,097 in the previous Chromy_Gen run to 24,516. The additional constraints added about 37,000 units back into the sample. Of these additional units, about 32,000 were in the American Indian frame. This brought the sample size to 2.3 million.

Summary of the 1997 Design

This completed the sample design work for the 1997 SMOBE. The final estimated sample size to achieve the goals for the 1997 design was 2,330,107 units. A portion of the remaining units in the budgeted 2.5 million was used for the Survey of Women-Owned Business Enterprises, a companion survey to SMOBE. The balance of the units was used to augment the NATIONAL frame sample. The 1997 SMOBE design has the following features:

(1) There are five frames -- ASIAN, BLACK, HISPANIC, AMERICAN INDIAN, and NATIONAL.

(2) Every sampling unit was classified by industry. The coarsest classification was the 2-digit SIC code. Some classifications were further sub-divided so that accurate estimates could be made at the 4-digit SIC level.

(3) Sampling units with receipts exceeding determined cutoffs were selected with certainty. This allowed for some control over the variation of the estimates of receipts, payroll and employment for each sampling stratum.

(4) If a minority-owned firm count estimate was greater than or equal to 100 for a frame-state-industry category in the ASIAN, HISPANIC, or AMERICAN INDIAN frames, the corresponding target coefficient of variation was 10 percent. If the estimate was less than 100, the corresponding target standard error was 10.

(5) All frame-state and frame-industry minority-owned firm count estimates from the ASIAN, HISPANIC, and AMERICAN INDIAN frames had a target coefficient of variation of 10 percent. Chromy_Gen software was used to minimize the sample size needed to simultaneously meet the frame-state-industry and frame-state and frame-industry constraints.

(6) CV constraints on firm count estimates of Asian, American Indian, and Hispanic- owned businesses found in the BLACK and NATIONAL frames were imposed to achieve coefficients of variation of 10 percent or less at the state and industry levels.

(7) No explicit measures were taken to ensure the accuracy of estimates for employers or for estimates by legal form of organization. A measure of variability control was imposed by sorting each sampling stratum by these characteristics before systematically selecting the sample.

Future Plans For the 2002 Design

Although we have decreased the sample from 7.0 to 2.5 million, we still have hopes of making the design even more efficient. Part of the solution will lie in our ability to accurately partition the universe of business enterprises into the various frames. Williams and Moore (1998) describe our initiatives in this area. The other part of the solution will be our ability to obtain maximum utilization of the information on our enhanced frames. In this section, we briefly describe our agenda.

Allocate More Efficiently. The current sample size determination algorithm assumes that the sample is drawn randomly. We are actually using systematic sampling to select the sample within each sampling cell, using a strategic sort based on the probabilities of correct frame assignment computed for each unit in the universe. This systematic selection is likely to be more efficient than simple random sampling. We may be able to use this design efficiency to reduce sample size.

Another approach we will investigate is substratifying the sampling strata by ranges of the probabilities of frame assignment, to improve sample allocation.

Improve the Reliability of Auxiliary Estimates. We may try imposing variability constraints on receipts, payroll, and employment. We may also try to use a ratio estimator for these characteristics. The ratio estimator would use the probabilities of correct frame assignment, which are available for each unit.

Incorporate Small Area Estimation Into the Design. We receive many requests for more detailed estimates than we publish. We are conducting a pilot study in the 1997 SMOBE on the use of 5-digit ZIP Codes to identify potential minority-owned businesses. Every business which operates in a ZIP Code where over 50 percent of the population is neither White nor Black was assigned to

a minority frame. Consequently, most businesses in Hawaii were assigned to the ASIAN frame. We expect to be able to produce accurate domain estimates at the 5-digit ZIP Code level for Hawaiian-based businesses. If our research shows that ZIP Codes are a good identifier for minority-owned businesses, they will probably be a dominant factor in the probability assignment algorithm. We are hopeful that this will improve domain estimation at fine geographic levels, even if the stratification is at the state level.

We may also research the use of small area estimation techniques to produce accurate estimates at finer levels than were sampled. Ratio estimation for small domains will also be considered.

Conclusion

For the 1997 survey, we have provided an efficient design which meets many of the users' requests with a "reasonable" sample size of about 2.5 million. However, we are not finished. We are conducting research which we hope will lead to an even more useful and efficient survey for 2002.

References

CHROMY, J. (1987). "Design Optimization with Multiple Objectives," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 194-199.

GLASSER, G. J. (1962). "On the Complete Coverage of Large Units in a Statistical Study," *Review of the International Statistical Institute (Vol mc 30:1)* , pp. 28-32.

SANDS, M.S. (1993). "Frame Creation for the Survey of Minority-Owned Business Enterprises and the Survey of Women-Owned Businesses," *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, pp. 775-780.

WILLIAMS, A.C. and Moore, R.A. (1998). "Using Administrative Data to Enhance the Sampling frame for the 1997 Survey of Minority-Owned Business Enterprises," *Proceedings of the Survey Research Methods Section, American Statistical Association* (to appear).

ZAYATZ, L. and SIGMAN, R. (1995). "CHROMY_GEN: General-Purpose Program for Multivariate Allocation of Stratified Samples Using Chromy's Algorithm," *Economic Statistical Methods Report Series ESM-9502*. U.S. Bureau of the Census.

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.