USING ADMINISTRATIVE DATA TO ENHANCE THE SAMPLING FRAME FOR THE 1997 SURVEY OF MINORITY-OWNED BUSINESS ENTERPRISES (SMOBE)

Anthony C. Williams and Richard A. Moore, Jr. , U.S. Bureau of the Census
Richard A. Moore, Jr. U.S. Bureau of the Census, Washington, DC 20233

Keywords: Minority-owned Business Enterprises; Domain Estimation; Administrative Data

I. Abstract. The 1997 Survey of Minority-Owned Business Enterprises (SMOBE) provides data on the number, receipts, payroll, and employment of minority-owned sole proprietorships, partnerships, and corporations. This sample survey is based on a design that consists of stratifying businesses by state, industry, and race of the owner(s). Employer and nonemployer business registers contain the state and industry codes. However, very little accurate data on the race of each owner is available. Thus, one obstacle to designing an efficient sample for this survey is to identify the small proportion of minority businesses prior to sampling. Information on the Census Bureau registers, in conjunction with information obtained from various administrative data sources, is used to assign the most likely race of the owner(s) to each business enterprise. This paper examines the sources of the administrative data, assesses the reliability of each source, and indicates the major logistical and methodological concerns encountered while creating the 1997 SMOBE sampling frame.

II. Purpose of the Economic Census. For the calendar years ending in 2 and 7, the U.S. Bureau of the Census conducts an Economic Census. Its primary goal is a complete enumeration and tabulation of all private non-farm business activities based in the United States. To achieve this for 1997, the Census Bureau produced two registers containing companies that operated during 1997: (1) the 1997 Standard Statistical Establishment List (SSEL) consisting of about 5.3 million employer companies, and (2) the 1997 Nonemployer Database (NEDB) consisting of about 13.1 million nonemployer companies. Each company record on these registers contains information about: (1) the various locations from which it operated, (2) its principal industrial activity, (3) its receipts, (4) its payroll, and (5) the number of people which it employed. This information allows the Census Bureau to produce detailed aggregate economic statistics for domains defined by both an industrial activity and a geographic area (e.g., restaurants in the Washington, DC Metropolitan Statistical Area (MSA)).

III. Purpose of SMOBE. As part of the Economic Census, SMOBE provides similar estimates restricted to companies having over 50 percent of the ownership being minority (Hispanic, Black, Asian-Pacific Islander, or American Indian-Alaskan Native). The 1997 SSEL and NEDB contain all the information necessary to produce these statistics except the predominant race of ownership for each company. SMOBE selects a representative sample of companies, from which it requests information about the race and ethnicity composition of the owners. Based on the 1997 responses of a sample of almost 2.5 million companies[1], SMOBE will publish cell estimates for each combination of (1) the predominant race of ownership of the company, (2) the company's state of operation, and (3) its primary 2-digit Standard Industrial Classification (SIC) code.

IV. Challenge Presented to SMOBE. SMOBE uses domain estimation to produce the 1997 minority-owned business estimates. The survey's big challenge is trying to make detailed estimates for relatively small domains. Table 1 provides counts of the number of companies in the U.S. by predominant race of ownership. As the table shows, the Asian-Pacific Islander (referred to as "Asian" in the remainder of this paper), Hispanic, and American Indian-Alaskan Native (referred to as "Am. Indian" in the remainder of this paper) ownership classifications each consist of less than 1.0 million firms.

Table 1. Total Number of Companies in the US By Race of Owner (1996)

Race	# Companies (Mil.)
White	16.0
Black	0.6
Asian-Pacific Islander	0.9
Hispanic	0.8
Am. Indian-Alaskan Native	0.1
Total	18.4

[1] In addition, the sample includes about 125,000 sole proprietorships that responded to the previous SMOBE. These cases are selected with certainty. Because each sole proprietor rarely changes his race and ethnicity, these cases are not mailed. Instead their 1992 SMOBE race and ethnicity responses are used to make estimates for the 1997 survey.

When cells are defined by industry classification and state detail, many contain very few companies. A sample large enough to make detailed estimates for all of these cells would be very expensive. Hence, SMOBE does not attempt to achieve a high degree of precision for all estimates. The goal of the 1997 SMOBE is to provide accurate (a coefficient of variation of 10 percent or less) firm count estimates for all standard publication cells for which the estimate is at least 100 companies. Although no formal variance constraints are placed on the auxiliary statistics (employment, payroll, and receipts), companies with extremely large receipts are sampled with certainty.

Table 1 is segmented into two groups: (1) White/Black, and (2) the other minorities. For the former group, accurate race-industry-state estimates can be made with a relatively small sample. White-owned businesses are so prevalent, that a only a small sample is required to produce accurate estimates. Over 0.5 million of the Black-owned businesses are sole proprietorships. For each sole proprietorship, we have the Social Security Number (SSN) of the owner. Administrative data, in particular the self-designated race on an individual's application for a SSN, fairly accurately identifies each Black sole proprietorship. We must canvass only a small proportion of partnerships and corporations to estimate the number of these which are Black-owned.

If we could pre-identify Asian, Hispanic, and Am. Indian sole proprietors as easily, we could also use a small sample to make detailed estimates for these three groups. Unfortunately, this is not the case. Prior to 1981, the self-designated race choices on the SSN application were White, Black, and Other. While a self-designation of "Other" is a strong indication that the owner belongs to one of these minorities, the converse is not true. People, in general, don't like to classify themselves as "Other". Consequently, many of these minorities restricted themselves to the choice of "White" or "Black." Prior to 1981, a high percentage chose "White" and the self-designated race on the SSN application lost all of its discriminating power. Beginning in 1981, race categories on the SSN application were expanded to include Asian, Hispanic, and Am. Indian. Hence, the SSN race does have some limited discrimination power for identifying businesses owned by these three minority groups.

V. Four-Step Approach for Identifying Potential Minority-Owned Firms. Since Asian-, Hispanic-, and Am. Indian-owned businesses are not readily identified by the race on the SSN application, SMOBE must make use of a wide variety of administrative data sources, to identify as many potential Asian, Hispanic, and Am. Indian-owned businesses as possible. Some of these sources accurately predict the race of ownership. Most, however, provide only a slight indication that the company may be minority-owned. The 1997 SMOBE uses a four-step approach in its frame construction and sample design scheme: (1) infer the race of ownership, (2) estimate the accuracy of this inference, (3) use the inference in the stratification, and (4) tailor the sample selection, weighting, and variance estimation to take advantage of the accuracy distribution. This section describes each of these steps in detail.

Step 1: Infer Ownership Race. Since the inception of SMOBE in 1969, the first step of the survey's frame construction has always been to infer a race of ownership to each company. With each successive SMOBE, we continue to bias this inference toward the difficult to identify minorities (Asian, Hispanic, and American Indian) by assigning companies to these minority frames even when only scant evidence is available to suggest minority ownership. For the 1997 survey, this is no different.

Some sources of data generally lead to accurate inferences. Below are 4 such examples.

(1) A business is found on a "Top 100" list (e.g., the Top 100 Hispanic-owned Businesses) in a nationally syndicated publication.

(2) A company gave race and ethnicity responses to a previous SMOBE. Although partnerships and corporations may change the composition of their ownership over the years, it is a safe assumption that the overall racial composition only occasionally changes.

(3) A company has a single owner. This person acquired his SSN after 1980 and chose a race other than White or Black.

(4) A company has a single owner, with an SSN self-designation of "White". The business is located in California and the owner's surname is "Lee". We have evidence that 8 of every 9 Lee's in California are of Asian descent.

Not all sources of administrative data lead to accurate inferences. Below are listed some sources of inaccurate inferences.

(1) A business has a single owner. This individual was born in Panama. The owner's country of birth is the only indication that he may be a minority. It is possible that the owner is the offspring of an American business person or a member of the U.S. armed forces.

(2) A business operates in an area which has a large percentage of minorities (e.g., near an Indian reservation, in a "Chinatown" district, or in a

Cuban suburb of Miami). This gives some indication that the business could be owned by a particular minority, but we can't estimate how often that the inference is accurate.

(3) A business has several owners. We know the SSN's of each. One is self-designated as "Asian" on his SSN application. The remainder of the owners self-classified themselves as "White". Although this provides some indication the business may be Asian-owned, it also raises several questions. (e.g., Does each own an equal share? Are some of the "White" owners really "Asian" who were restricted to the White/Black/Other classification choice?)

(4) A company has a single owner, self-designated as "White" on his SSN application. The business is located in Maine and the only indication of minority ownership is the owner's surname, Lee. We have evidence that only 1 of every 40 Lee's in Maine are of Asian descent.

Step 2: Estimate the Accuracy of the Race Inference. There are numerous sources from which we can infer a company's race of ownership. Some of these sources provide extremely accurate inferences, while others lead to incorrect inferences a high percentage of the time. Now we must develop a method to differentiate accurate and inaccurate inferences.

For the 1997 SMOBE, we used the set of 0.9 million firms that responded to the 1992 SMOBE (the most recent race data available) to determine the proportion of times that the inference was correct. We found such things as:

(1) In 1992, there were 2,243 sole proprietorships whose owner was named "Smith" in Oklahoma. Of these owners, 20 were American Indians. Hence, the accuracy for inferring a 1997 sole proprietor named "Smith" as Am. Indian is 0.009 (20/2243).

(2) There were about 222,000 sole proprietors indicating SSN race classification "Other". The 1992 SMOBE estimated that about 174,000 of these are Asian-owned. Hence, we assume that about 78.2 percent of all 1997 sole proprietorships exhibiting this characteristic are Asian-owned.

These and similar results were used when assigning measures of accuracy.

Before proceeding to Step 3, let's review our progress towards attaining the goals of the initial steps. Because companies assigned to one of these races are more heavily sampled and minority-owned estimates will have lower variability , the goal of Step 1 was to infer the ownership race of Asian, Hispanic, or Am. Indian to as many businesses as possible. Table 2 shows that we have inferred one of these races to about 11.1 million (or 60 percent) of the companies. Table 2 also shows that, based on responses to the 1992 SMOBE, we infer 0.3 million businesses as Black-owned and 0.1 million as White-owned.

The final row of the table shows that there are 6.9 million companies for which this administrative data gives us no indication of the race. Results from the 1992 SMOBE estimate that about 1.5 percent (or 0.1 million) of these are actually Asian-, Hispanic-, or Am. Indian-owned. Hence, we have inferred a minority race to about 1.7 million (or 94 percent) of the targeted population.

Table 2. 1997 SMOBE Race Inference Results

	Millions of Companies	
Inference	# Inferences	# Asian, Hisp, Am Indian
Hispanic	5.2	0.8
Asian	5.0	0.8
Am. Indian	0.9	0.1
Sub Total	11.1	1.7
Black	0.3	
White	0.1	
No Indication	6.9	0.1
Total	18.4	1.8

Although we have inferred minority-owned status to a large number of companies, many of our inferences are incorrect. The goal of Step 2 was to develop a measure of accuracy which segregates accurate from the inaccurate inferences. By construction, the accuracy of each inference is now a continuous variable between 0.000 and 1.000. Inferences with accuracies near 0.000 are often incorrect, while those with inferences near 1.000 are usually correct. Hence the accuracy can be used to discriminate accurate and inaccurate inferences. Table 3 indicates that distribution of the accuracies are reasonably bimodal. It shows that about 8.89 million (or 80 percent of all) inferences are estimated to be incorrect at least 7 out of 8 times, while about 0.96 million (or about 9 percent of all) inferences are estimated to be correct at least 7 out of 8 times. Only about 1.26 million (or 11 percent of all) inferences fall in a relatively large "gray" area.

Table 3 also illustrates the importance of making inferences for the . The 8.89 million weakest inferences contain about 0.24 million minority-owned businesses. If these inferences were not made, these cases would appear on the "No Indication" row of Table 2.

Table 3. Distribution of the Race Inference By Accuracy Range

Accuracy Range	Inferences (Million)	Correct Inferences (Million)	Average Accuracy
0.000 - 0.125	8.89	0.24	0.027
0.126 - 0.875	1.26	0.56	0.444
0.876 - 1.000	0.96	0.90	0.938

Step 3: Use the Race Inference in the Stratification and Sample Design. The overall goal of the survey is to make accurate state by industry estimates for each minority group. The geographic and industrial information on the 1997 SSEL and NEDB categorizes all businesses by state and industry cell. The race inferences allow us to identify the companies which are more likely to be owned by a particular minority. The next logical step is (1) to stratify all businesses by inferred race, state, and industry code, and then (2) select a sample from each stratum. Consider the following simple hypothetical example.

Hypothetical Example. We want to estimate the number of Am. Indian-owned widget producers in Oklahoma with a standard error of 10 percent of the estimate. The chart below shows that we have identified 100 potential such businesses. In this simple example, all race inferences were based on the owner's surname. For 10 of these firms, we are very certain that most of the inferences are accurate, because the owner is named "Runningbear". We have little confidence the other inferences, because each owner is named "Smith".

Owner's Name	# Inferences	Accuracy
Runningbear	10	0.991
Smith	90	0.009

Stratified Simple Random Sampling. Continuing in this vein, the most primitive sample design is a stratified simple random sample. In this case, one could estimate the proportion of Am. Indian-owned businesses in the stratum by averaging the accuracies. This would yield

$$\hat{p} = 0.1072 .$$

Even with a finite population correction, one would have to randomly select 90 out of the 100 companies to get the desired precision. Why such a large sample? The sample must be representative of the units in the stratum. Any representative sample will contain about 1 Runningbear for every 9 Smith's. A sample of 90 businesses is necessary to "guarantee" that this ratio is approximately 1 to 9. Let's look for a more efficient sample design.

Step 4: Tailor the Sample Design To Take Advantage of the Bimodality of the Accuracy Distribution. In general, stratified simple random sampling leads to exorbitant sampling rates. Let's consider some simple alternatives.

Further Stratify by Accuracy and Simple Random Sample. The accuracy measure allows us to separate the accurate inferences and the inaccurate inferences into different "classes". Having done this, we can use smaller samples to make relatively accurate estimates for the number of minorities in each class and then add the two estimates together. In fact, if we stratify by accuracy in this example, we would only have to sample 44 businesses ---(4 Runningbears and 40 Smiths) --- to attain our desired precision.

Stratified Cluster Sample. Another approach would be to recognize that any representative sample contains 1 Runningbear for every 9 Smiths. From this, we can form 10 mutually exclusive clusters, each cluster containing 1 Runningbear and 9 Smiths. Based on the information provided, it is likely that all of these clusters would contain at least one Am. Indian-owned business; and possibly 1 of the 10 clusters would contain 2 Am. Indian-owned firms. Typically, the set of the estimates for the 10 clusters would be {2, 1, 1, 1, 1, 1, 1, 1, 1, 1}. Based on the variation of the cluster estimates, one would estimate that there are 11.00 American-owned firms in this stratum with a standard error of 1.00."

The measure of accuracy allows us to form "almost identical" clusters, estimate at the cluster level, then evaluate the stability of the aggregate estimate. More formally stated, we could cluster, select a sample of clusters, and enumerate all units within the cluster. To attain our desired precision for the Runningbear-Smith problem, we would have to select 5 clusters and enumerate 50 firms total.

Ratio Estimation. A fourth approach would be to use the ratio estimator. This allows for any sample design. Assume the accuracies are truly unbiased, then let

$$R = \frac{\textit{Estimated No. Minorities in Stratum}}{\textit{Estimated No. Minorities in Sample}}$$

$$= \frac{\sum_{i \text{ in Stratum}} Accuracy_i}{\sum_{j \text{ in Sample}} Accuracy_j} .$$

By weighting each response by R instead of "N/n", you can get a reasonably accurate estimate, even when the sample is not representative.

Although we are not planning on incorporating the ratio estimator into the sample design, we are investigating the feasibility of its use for such projects as:

(1) adjusting weights to compensate for non-response,
(2) providing estimates for the auxiliary strata statistics (employment, payroll, receipts), and
(3) providing some sub-state (e.g., county or MSA) and more-detailed (e.g., 4-digit SIC) industry level estimates.

VI. Conclusion. Since its inception in 1969, SMOBE has always used administrative data to enhance the sampling frame by inferring a race of ownership to each company. We have learned that a large percentage of the administrative data which we receive is not as "clean" as we would like. Some of the more frequent problems are listed below.

(1) Administrative sources contain incorrect information (e.g., the information is several years old) about a large number of companies. Although our race inferences are logical, many are incorrect because our premises are incorrect.
(2) Even if the information on a company is accurate, it is sometimes difficult to correctly link to the company on the SMOBE frame. Often name and address are the only sources of linkage. Companies tend to use non-standard abbreviations and frequently add strimgs to or delete them from their names (e.g., a company may appear as ABC ASSOC LTD on one register and as A B CARR & ASSOCIATES on the 1997 SSEL).
(3) Accurate administrative information doesn't always lead to a positive identification of a minority-owned firm. For example, only 1 in 40 sole proprietors in Maine named "Lee" are Asian. Inferring a race of Asian to all "Lee" sole proprietorships leads to a large number of incorrect inferences.

In constructing the 1997 SMOBE sampling frame, we learned that "dirty" administrative information is better than nothing at all, provided

(1) we can distinguish the accurate inferences (i.e, those made from "clean" information) from the inaccurate ones;
(2) we can quantify the accuracy of each inference, using a sample of cases where the correct value of the inference is already known; and
(3) we can incorporate both the inference and the accuracy into the sample design.

In extreme cases, such as the Runningbear-Smith example, our practice of using "dirty" administrative records can lead to drastic cuts in the required sample size without sacrificing precision.

VII. Future Improvements. There have been major changes to the 1997 SMOBE. However, more changes may be possible in the future. Some aspirations for future SMOBEs include:

(1) Automate the Record Linkage Procedure. We identified several lists of minority-owned companies, which contained over 100,000 companies each. Unfortunately, we did not use these lists, because our 1997 record linkage procedure was manual. For future SMOBEs, we hope to automate this procedure by using software developed by Winkler (1993).
(2) Use Step-wise Discrimination. The 1997 frame contains 18.4 million companies. We currently gather administrative information from 14 different sources. This results in over 250 million pieces of information. As the availability of access to information increases (e.g. the Internet), we anticipate the number of sources to increase. Many of these sources will provide duplicate or less reliable information. Because we cannot afford to carry 500 million pieces of information throughout the frame construction and sample selection procedures, we plan to use step-wise discrimination software at the beginning of the procedure to eliminate all information except that essential for segregating minority-owned from non-minority-owned companies.
(3) Use a Multivariate Approach (e.g., Logistic Regression) to Determine Accuracies. In 1997, we used a hierarchy of sources to infer the race: (a) Do we have a response from a previous SMOBE for the company? (b) If not, is it found on a "Top 100" list? (c) If not, what is the owner's race on his SSN application?, etc.

For future SMOBEs, we would like to evaluate all relevant information simultaneously. Subtle differences in operations may make huge differences in how we measure the accuracy of our inferences. For example, suppose we have two very similar sole proprietorships. Both are operated by a Mr. Lee, who lives in Maine. Both operate from the same 5-digit ZIP Code. Business #1 is restaurant (SIC 5812), while Business #2 is a bar (SIC 5813). Is it possible Business #1 is probably Asian-owned, while Business #2 is probably not? Is it also possible that multivariate techniques may produce more bimodal distributions of accuracies?

(4) Use Variance Replication Software to Prove the Stability of the Measures of Accuracy. The distribution of the accuracies within a given stratum is an essential component to the calculation of the minimum sample size necessary to satisfy the variance constraints. Although our 1997 accuracy measures may be unbiased, we have yet to examine their stability. Basically, we need to show that

$$\hat{X}_{Stratum} = \sum_{i \text{ in } Stratum}^{N} Accuracy_i \approx No.\ Minorities\ in\ Stratum,\ and$$

$$\hat{X}_{Sample} = \sum_{j \text{ in } Sample}^{n} Accuracy_j \approx No.\ Minorities\ in\ Sample,$$

for any arbitrary stratum and most arbitrary subsets of this stratum. We plan to accomplish this with VPLX software, developed by Fay of the U.S. Bureau of the Census.

(5) Develop an Optimal Sample Design. Assume the multivariate methodology produces an unbiased measure of accuracy for each race inference and that the distribution of these accuracies is bimodal within each stratum. We are now in a position to determine which sampling procedure (simple random sampling after further stratifying by accuracy, stratified cluster sampling, simple random sampling using a ratio estimator, etc.) provides high precision estimates for sample sizes that meet our budget constraints. After choosing a new sample design, we must develop the accompanying model-based variance calculation procedure.

(6) Use Optimal Allocation Software. SMOBE provides estimates for nested domains (e.g., estimates at the 4-digit levels which sum to estimates at the 3-digit SIC levels ...). In some cases, the variance constraints differ at the various levels. To provide the desired precision at each level, the 1997 SMOBE used optimal allocation software for the first time. The software was developed by Zayatz and Sigman (1995).

VIII. Ultimate Goal for the 2002 SMOBE. For the 1997 survey, the goal was to accurately (i.e, with a coefficient of variation of less than 10 percent) estimate the firm count for all race by state by 2-digit SIC domains containing over 100 companies. For the 2002 survey, let's assume that we successfully implement many of the 6 projects mentioned in Section VII. Let's also assume that the budget allows us to sample about 2.5 million companies. To what goal can we reasonably aspire? After the sample is selected and the race and ethnicity responses collected, we hope to have the ability to produce (upon demand) accurate firm count, employment, payroll, and receipts estimates for any domain (regardless of the ethnicity, geographic, or industrial level requested), provided that domain contains at least 100 companies. For example, suppose the Dallas, TX MSA contains 300 Mexican-owned restaurants. Even though the 2002 SMOBE will be designed to produce accurate estimates for Hispanic-owned restaurants and bars in Texas, we hope to provide accurate firm count and auxiliary variable estimates for Mexican-owned restaurants in the Dallas MSA.

IX. Bibliography

Moore, Richard , Carol Caldwell, and Ruth Detlefsen (1998), "Changing the Sample Design to Meet User Needs --- The Survey of Minority-Owned Business Enterprises --- Past, Present, and Future," Proceedings of the Section of Survey Research Methods, (to appear).

Winkler, William (1993), "Matching and Record Linkage,"in Business Survey Methods (Brenda G. Cox, editor), New York: John Wiley and Sons, pp 355-384.

Zayatz, Laura and Richard Sigman, (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.

X. 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.