THE EVOLUTION OF SURVEY METHODOLOGY FOR A COMPANY SURVEY OF CAPITAL EXPENDITURES

Marie C. Stetser, John Michael Goodloe, Mark S. Sands, U. S. Bureau of the Census
John Michael Goodloe, Company Statistics Division, U.S. Bureau of the Census, Washington, DC 20233

Keywords: Stratification, Poststratification, Sampling Frame, ACES

I. INTRODUCTION

The Annual Capital Expenditures Survey (ACES) is a company based survey that collects data on fixed assets and depreciation, expenditures for new and used structures and equipment, expenditures by industry category for the industries in which the company participates, total sales and receipts, as well as sales and receipts data for the three largest industries of the company. Data are collected for the year prior to the year in which the survey is conducted. The current 1997 ACES is based on a probability sample of approximately 46,000 companies. The first ACES was for data year 1992.

The survey was conceived as the principal source of comprehensive statistics on business investment in buildings and other structures, machinery, and equipment for the domestic nonfarm U.S. economy. These data are important to economic policymakers concerned with the adequacy of investment in plant and equipment in the U.S. economy. Prior to the ACES, data on capital investment came from a variety of sources, none of which covered the entire economy from the standpoint of a probability-based sample. The data are also useful in the evaluation of productivity, competitiveness of the U.S. economy, changes in industrial capacity, and overall economic performance.

ACES is a source of capital expenditures data for other agencies within the U.S. Government. Among these are the Bureau of Economic Analysis, the Bureau of Labor Statistics, and the Federal Reserve Board. The data are also available to private sector businesses and researchers interested in capital investment.

II. SCOPE OF PAPER
A. General Issues

This paper addresses specific issues and challenges of an enterprise- or company-level[1] survey of capital expenditures. Some issues arise because the sampling unit for ACES is an enterprise while the list from which the sampling frame is constructed is establishment-based. This is critical because it is

---

[1] The terms *company* and *enterprise* are used interchangeably.

important that all establishments be accounted for in order to insure coverage of the entire economy. Another issue arises due to problematic industry classification of multi-establishment enterprises that have activity in more than one industry. Accurate industry classification for companies is needed for efficient sampling. There is another problem related to collecting capital expenditure data. Capital expenditures are a volatile component in a dynamic, changing economy. Companies often make substantial capital expenditures before they are fully operational. Insuring that these enterprises are included in the sampling frame is a challenge.

B. Population of Interest

The population of interest for ACES includes all domestic, nonfarm businesses located in the 50 states and the District of Columbia with the potential for capital investment. Government-owned operations and foreign operations of domestic businesses are excluded. ACES is a company- or enterprise-based survey rather than an establishment-based survey. An establishment is defined as "an economic unit, generally at a single physical location, where business is conducted or where services or industrial operations are performed." An enterprise is defined as "an economic unit comprising one or more establishments under common ownership or control" (Walker, 1997). Enterprises which consist of two or more establishments are commonly referred to as multiunits; enterprises which consist of only one establishment are commonly referred to as single units. The enterprise was chosen as the sampling unit for ACES in order to maximize coverage of the capital expenditures by industry while maintaining a manageable and affordable sample size (Champion, 1991). Moreover, information on capital expenditures often will be more readily available and will contain more detailed information at the enterprise level than at the establishment level.

III. SPECIFIC ISSUES
A. Sampling Frame
1. Business Registers

ACES creates its frame from the Census Bureau's primary business register, the Standard Statistical Establishment List (SSEL), as well as an auxiliary register, the Nonemployer Database (NED). The SSEL serves as a list frame for the economic censuses and most of the economic surveys at the Census Bureau. The primary source of data for the SSEL is the

Business Master File (BMF) provided by the Internal Revenue Service (IRS). The BMF provides names, addresses, and other identifying information of all businesses which are potential tax paying businesses. Additional administrative data for the SSEL comes from quarterly payroll tax returns and annual income tax returns. Employer businesses report data to the IRS by the Employer Identification Number (EIN) which is assigned by the IRS; hence, the IRS provides administrative data to Census by EIN. Often the EIN corresponds to a particular establishment on the SSEL, but this is not always the case. Multiunit businesses may have several establishments reporting under one EIN. Other sources of data for the SSEL include administrative data from the Social Security Administration (SSA) and the Bureau of Labor Statistics (BLS), information collected by the Census Bureau's Company Organization Survey (COS), information from the quinquennial economic censuses, and additional updates from current economic surveys.

The auxiliary NED is somewhat different from the SSEL in that it is reconstructed annually rather than simply updated continuously. The primary source of data for nonemployer companies is annual income tax returns. Nonemployer corporations and partnerships are identified by EIN, but nonemployer sole proprietorships are identified by the social security number of the owner. For a given year, income tax returns are compiled until about December of the following year. At that time, the nonemployer database for the given year is complete, and the processing begins again for the next tax year.

The target population for a survey that collects information on sales and receipts or payroll, for instance, may be different from that of a survey of capital expenditures. This is especially true when defining if a new business, or birth, is eligible for selection in a survey. For many economic surveys, a business becomes a birth and is added to the business register when IRS payroll information is first reported. Employer businesses are first identified when a business applies for an EIN and begins reporting quarterly payroll information. Nonemployer businesses are first identified when businesses report year-end sales and receipts data on year-end tax returns. This system of birth identification is reasonable for most surveys collecting sales information, but is not adequate when collecting capital expenditures information. As part of normal start-up, new businesses often begin making capital expenditures before they have payroll or receipts. For this reason, coverage of new business investment activity for ACES presents an additional challenge.

2. Frame Construction

The ACES universe is generally divided into two segments, the ACE-1 segment and the ACE-2 segment.

The ACE-1 segment consists of companies with one or more employees (prior to the 1996 ACES, the ACE-1 segment consisted of companies with 5 or more employees). ACE-1 companies are mailed a questionnaire that asks for capital investment data by industry; hence, data for the ACE-1 segment are published at industry level detail. The ACE-2 segment was first introduced for the 1993 survey with the intention of only collecting data from them every two years. Due to data users' needs, data for ACE-2 companies have been collected annually since 1995. The ACE-2 segment consists of nonemployer companies (prior to the 1996 ACES, companies with 1 to 4 employees were part of the ACE-2 segment). Data for the ACE-2 segment are collected and published at the total level only.

a. ACE-1 Sampling Units. Sampling units for the ACE-1 segment of the frame are constructed as follows. The first step is to extract from the SSEL all of the establishments which are in scope for ACES by census file number (CFN). Prior year payroll, prior year employment, and a 4-digit Standard Industrial Classification (SIC) code are also extracted for each establishment. Prior year payroll and employment data are used in the stratification process described in later sections.

The second step is to consolidate multiunit establishment records into company records. The SSEL contains information that identifies establishments belonging to a particular enterprise. The Census Bureau obtains this information via the annual Company Organization Survey and the quinquennial economic censuses (Walker, 1997), as well as from updates received from periodic surveys. Establishments which are part of multiunit companies and are in scope of the ACES survey are grouped by company. Payroll and employment for in-scope establishments are summed to create company-level payroll and employment.

The third step is to assign ACES industry codes to each company. Each establishment on the SSEL contains a 4-digit SIC code. For purposes of ACES stratification and tabulation, the SIC is converted into a 3-digit ACES industry code. For each year of the production survey prior to 1997, there were 94 different ACES industry codes, including a category for unclassified companies. In 1997, three additional codes were added to provide greater detail for health service industries. Classification of single-unit companies is a simple SIC to ACES industry code conversion. The assignment of ACES industry codes is not so simple for multiunits. Because an SIC exists at the establishment level, a particular enterprise might participate in several different types of industrial activities. Before assigning an ACES industry code to a multiunit company,

establishment payroll is tabulated by trade area (i.e., mining, manufacturing, retail, etc.) and then by 4-digit SIC for each company record. The company is assigned to the trade area with the highest total payroll. Within that trade area, the company is assigned to the SIC with the highest total payroll. For each company, the SIC is converted to the corresponding ACES industry code. The final 1997 ACE-1 segment of the universe resulted in over 4.7 million single and multiunit company records.

b. ACE-2 Sampling Units. The ACE-2 segment is partially composed of records from the SSEL and partially of records from the NED. Sampling units from the SSEL are constructed in the same manner and at the same time as those for the ACE-1 segment. Those cases with zero prior year employees are placed in the ACE-2 segment.

Sampling units from the NED are simply the records as they exist on this file. The sampling units are assigned to different substrata as detailed later in this section.

Because data sources for the NED are different from those for the SSEL, an unduplication process is necessary between the two frames. Corporations and partnerships are matched to the SSEL frame on EIN wherever possible and unduplicated. Beginning in 1997, sole proprietorships were matched to the SSEL frame on social security number wherever possible and unduplicated.

Table 1 gives a summary of substratification of the ACE-2 companies.

Companies from the SSEL not included in the ACE-1 segment are divided into three substrata. Companies with zero prior year employment and positive prior year payroll are included in substratum B. The employment data field on the SSEL refers to employment for the pay period including March 12. Substratum B includes seasonal companies and companies going into business after March 12 or out of business before March 12. Substratum B consisted of approximately 728,000 companies for the 1997 ACES.

SSEL companies with zero prior year employment and zero prior year payroll are included in substratum C. Substratum C is a mixture of companies which do not fall in any other category. Substratum C consisted of about 2.3 million companies for the 1997 ACES. Many of these companies are new companies that may have current year payroll reported, but not prior year payroll. Other companies in substratum C have earlier payroll but have not yet been purged from the SSEL (companies remain on the SSEL for nine quarters after payroll reporting has been discontinued). Still other companies in this substratum have some administrative data such as receipts or interest income reported within the past two years.

Table 1: Summary of ACE-2 Substrata		
Sub-stratum	Source	Description
A*	SSEL	Companies with 1-4 employees in the prior year. These companies were moved to the ACE-1 segment starting in 1996.
B	SSEL	Companies with zero employees in the prior year, but prior year payroll > 0.
C	SSEL	Companies with zero employees in the prior year and zero prior year payroll.
D	NED	Nonemployer partnerships and corporations.
E	NED	Nonemployer sole proprietorships.
F	SSEL	Companies with EINs which are "F941 active" but have no administrative data reported. Included for the first time in 1997.

*Prior to 1996, SSEL companies with 1 to 4 prior year employees were included in substratum A of the ACE-2 stratum. Beginning in 1996, companies with 1 to 4 prior year employees were included in the ACE-1 segment. Substratum A is thus no longer part of the ACE-2 segment.

Questions have been raised as to why the ACES frame continues to include companies which seem to have gone out of business (i.e., substratum C). A significant portion of these companies are determined to be out of scope over the course of the survey. A significant portion never respond at all to the survey, leaving the scope of the business unknown. Over the past year, substratum C has been examined more closely to determine if there are portions of the substratum which could be eliminated using the administrative data available at the time of frame creation.

The administrative data on this part of the 1996 frame was examined closely. First, businesses with prior or current year payroll were extracted. Then, of the remaining records, businesses with current or prior year administrative data such as receipts, assets, interest income, or rent were extracted. Of those records remaining, businesses with two years prior payroll were extracted; then the remaining records were grouped together. Table 2 shows the percentage of the substratum frame in each category.

The 1996 sample records were then matched back to the corresponding portion of the frame. Total capital expenditures for each category were examined. Table 2 shows the percentage of the substratum estimate representing that portion of the substratum frame.

Table 2: Characteristics of Substratum C Administrative Data		
	Percentage of 1996 Frame Represented	Percentage of 1996 Substratum Estimate*
Current payroll > 0	31.4	58.9
Prior or current administrative data > 0	36.7	15.8
Two years prior payroll > 0	24.2	9.5
None of the above	7.8	15.8

* Refers to the post-stratified substratum estimate

While many surveys could justify excluding records which have no prior or current year payroll, this is not the case for ACES. Companies with current payroll do account for nearly sixty percent of the estimate for this substratum, but the remaining forty percent cannot be ignored. Obviously, there is capital expenditure activity in those categories. For this reason, the 1997 frame continues to include companies on the SSEL regardless of presence of payroll.

ACE-2 companies extracted from the NED are companies that have reported sales or receipts data on an annual tax return but have not reported payroll or employment information. Nonemployer records are divided into two substrata, one for corporations and partnerships (substratum D) and one for sole proprietorships (substratum E).

Further examination of the 1996 frame revealed another potential source of capital expenditures which was previously missed. In 1997, about 1.4 million records exist on the SSEL where companies have an EIN, but have not yet reported any administrative data. This group of records is a potential source of births for ACES. The limited data for these records was examined. Of those 1.4 million records, approximately 710,000 were determined to be "F941 Active," meaning that the companies were expected to begin reporting quarterly payroll to the IRS. For the 1997 ACES, these companies became a new substratum (F); a random sample of 2,400 companies was selected from substratum F and mailed an ACE-2 questionnaire. The results of this mailing are pending the completion of the data collection for the 1997 ACES.

B. Stratification for Sampling

The ACES sample is partitioned into three main strata. The first stratum is a certainty stratum consisting of the ACE-1 companies with the largest employment. The second stratum consists of the remaining ACE-1 companies. The second stratum is further substratified by size of payroll within ACE industry code. The size strata will be described in more detail below. The third stratum contains all of the ACE-2 companies and is further substratified according to source (SSEL or NED) and other pertinent characteristics (number of employees, payroll, type of organization, etc.) as described in the previous section. Note that the ACE-2 segment is not stratified by ACES industry. SIC codes are not available for a large percentage of these sampling units. In addition, the sample sizes required for industry detail of this segment would be cost prohibitive.

As noted above, stratum 2 is further substratified by payroll. The payroll substrata are determined using the Lavallée-Hidiroglou (L-H) method (Lavallée and Hidiroglou, 1988; Slanta and Krenzke, 1996). The L-H methodology is an optimizing, iterative algorithmic approach which selects substratum boundaries. It is used in ACES to minimize sample size subject to a fixed coefficient of variation (CV) constraint. Payroll is used as a proxy for the true variable whose variance is to be minimized, capital expenditures. The L-H program, as applied to ACES, determines payroll boundaries that minimize sample size within each industry (note: the L-H method can also minimize the CV for a fixed sample size, depending on how it is used).

Prior to the 1996 survey, the second stratum was divided into three substrata within each industry using the L-H method. For the 1996 survey, the L-H method was used as before, but an additional substratum containing companies with from 1 to 4 employees was included in stratum 2 and was substratified on industry but was not processed through the L-H programs. This substratum was maintained to permit bridging between 1995 and 1996 survey estimates. For the 1997 survey, four substrata within the ACES industries were maintained but the separate 1 to 4 employee substratum was not kept. Instead all were payroll substrata. Programming resource constraints confined us to no more than four substrata.

We needed to determine the optimal number of substrata to use when including the companies with from 1 to 4 employees with the rest of the stratum 2 companies for the 1997 ACES and beyond.

The question was ultimately treated empirically because there was little available that would answer the question on a purely theoretical basis. Cochran shows that increasing the number of strata will decrease the variance (Cochran, 1977, p.132). However, if stratification is based on a different variable than the one that is being measured, the effect of adding more strata is diminished. If the correlation coefficient between the two variables, the one being measured and the one used for sampling, is of the order of .85 then the practical limit for the number of strata appears to be 6 (Cochran, 1977, p.

133).

The most recent sampling frame available at the time the empirical test was performed was the 1996 sampling frame. Therefore, the 1996 sampling frame was used for the test, with one modification: The companies with one to four employees were treated as a separate stratum for 1996; for the purposes of the test, these companies were added to all other noncertainty employer companies.

As indicated above, the L-H methodology assumes a fixed number of strata and a stratification variable that can be approximated by a continuous density function. The latter condition is fulfilled for ACES by payroll. The program then adjusts strata boundaries until sample size is minimized. The program was run on the noncertainty employer companies within the 94 industry strata assuming 3, 4, and 5 substrata and the resulting sample sizes were compared. We did not run the test for 6 substrata because we knew that programming and processing constraints would prohibit the implementation of this number. We did not test 2 substrata because it was clear that increasing the number of strata was producing the desired result so further testing in the other direction was regarded as unnecessary.

The final sample size results contain an additional adjustment. The components of the adjustment are a minimum sample size for an industry stratum and a minimum sample size for any substratum of an industry. This is referred to as an "s/r" rule, where "s" is the minimum sample size for the total and "r" is the minimum sample size for the substratum. After the L-H program is run, the resulting sample sizes are adjusted in accordance with the appropriate s/r rule. For the 3 strata test a 50/10 adjustment rule was used. The 4 strata test was run twice, once using a 50/10 adjustment rule and a second time using a 75/15 rule. The 5 strata test used a 75/10 adjustment rule.

Table 3 shows the results from the tests in terms of sample size. "Certainty Cases" refers to the cells where the L-H method along with our s/r rule resulted in every company within that cell being included in the sample. The number of certainty cases is an important consideration if increasing the number of strata causes the number of certainty cases to increase substantially. It is desirable to reduce the number of certainty cases in order to prevent these smaller companies from being sampled every year.

These results show clearly that increasing the number of strata results in a substantial decrease in the sample size when the target variance is fixed. Thus, increasing the number of strata from 3 to 4 resulted in a decrease in the sample size of about 6,000 cases. Increasing the number of strata from 4 to 5 reduced the sample size by another 3,000 cases. Increasing the

number of strata even results in a decrease in the number of certainty cases.

Table 3: Sample Sizes for Strata Tests		
Strata/Min	Sample Size	Certainty Cases
Actual 1996	19,189*	1,148*
3/10	19,302	1,268
4/10	11,877	649
4/15	13,063	771
5/10	9,838	523

*includes sample from separate stratum with 1 to 4 employees. These cases were included in all tests but not as a separate stratum.

The results of these tests clearly indicate that the degree of stratification has a substantial effect on sample size for given target levels of variance. Thus, the efficiency of the survey estimates relative to the size of the sample can be affected substantially by considering the impact of the number of strata for a given survey.

C. Poststratification

For the 1993 ACES, all companies that were on the SSEL with zero employees were included in one nonemployer substratum. Post-survey analysis of respondent data within this substratum showed that some companies with zero employees and zero payroll actually responded with capital expenditures. So, for the 1995 ACES, the substratum was split to place companies with zero employees but nonzero payroll in substratum B, and companies with both zero employees and zero payroll in a separate substratum C.

A review of 1995 ACES showed that many cases in substratum C responded with zero or very small capital expenditures. However, several companies responded with relatively large (and in some cases very large) capital expenditures. The decision was made to examine the characteristics of the responding cases using the most recent employment and payroll data.

This analysis showed that companies which had at least $50,000 in payroll were more likely to have the larger reported capital expenditures. So the decision was made to poststratify substratum C into two strata based on current year payroll.

To perform the poststratification, the original substratum C frame was matched to the more recent SSEL to obtain the administrative data for the survey year (instead of the prior year data used when the frame was constructed). Then the universe size and sample sizes were determined based on the poststratification criteria (e.g., for 1995, fewer than $50,000 in payroll, and $50,000 or more in payroll). Then the original sample

weights were adjusted based on the poststratification universe sizes and sample sizes.

This poststratification resulted in a lower variance on the estimated total and a more accurate estimate. This same procedure was used in the 1996 ACES survey for substratum C, except that the stratum was poststratified into three strata. Table 4 shows the percent decrease in substratum C and total ACE-2 segment variance due to poststratification when compared to the original stratification for substratum C.

Table 4: Reduction in ACE-2 Segment Variance due to Poststratification of Substratum C		
Survey Year	Reduction in Variance in Substratum C	Reduction in Variance in ACE-2 Segment
1995	63.9%	42.8%
1996	37.7%	15.7%

IV. SUMMARY AND CONCLUSIONS

This paper has discussed how ACES has attempted to deal with the issue of births both during frame creation and by poststratification. We have also discussed other ways that ACES has attempted to insure more complete coverage of the universe through better utilization of the sampling frame. We have discussed ways that ACES has made the sampling more efficient. For example, we have discussed how stratification was used to create a more efficient allocation of the available sample.

The Annual Capital Expenditures Survey is a data collection instrument that targets a rapidly changing business environment. The methodology that supports that instrument has evolved as we have learned more about the target population and the requirements of the data users. Significant progress has been made both in our understanding of the data we are collecting and in our ability to provide more precise and reliable estimates of the variables we are trying to measure. The tasks and requirements will continue to evolve in this interesting and challenging area. The methodology will continue to evolve as well.

References

Champion, Elinor J. and Berry, John H. (1991), "Annual Capital Expenditures Survey," paper presented to the Census Advisory Committees of the American Economic Association and the American Marketing Association, at the Joint Advisory Committee Meeting, April 25-26, 1991.

Cochran, William G. (1977), *Sampling Techniques*, 3rd edition, New York: John Wiley & Sons.

Lavallée, Pierre and Hidiroglou, Michel A. (1988), "On the Stratification of Skewed Populations." *Survey Methodology*, Volume 14, Number 1, pp. 33-43.

Slanta, John G. and Krenzke, Thomas R. (1996), "Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey." *Survey Methodology*, Volume 22, Number 1, pp. 65-75.

U.S. Bureau of the Census (1998), *Annual Capital Expenditures: 1996*, ACE/96, U.S. Government Printing Office, Washington, DC.

Walker, Edward (1997), "The Census Bureau's Business Register: Basic Features and Quality Issues," *1997 Proceedings of the American Statistical Association.*