IMPROVING THE DESIGN OF SAMPLES USED FOR BUSINESS SURVEYS CONDUCTED BY THE UNITED STATES BUREAU OF THE CENSUS[1]

David L. Kinyon, Donna Glassbrenner, Jock Black, and Ruth E. Detlefsen, U.S. Bureau of the Census
David L. Kinyon, U. S. Bureau of the Census, SSSD, Rm. 2651-3, Washington, DC 20233

Key Words: Sample design, Stratification, Sample sizes, Multiple reliability constraints

1. Introduction

The United States Bureau of the Census conducts monthly and annual surveys of the retail and wholesale trade areas and annual surveys of the service trade area to measure totals and trends relevant to the economy, primarily those pertaining to sales and inventories. The samples used for these surveys are periodically redesigned using data from the most recent Economic Census, the Company Organization Survey (COS), and administrative sources. For the most recent sample revision, the 2000 Business Sample Revision (BSR-2K), we accelerated our schedule to expedite replacement of the 1987 Standard Industrial Classification (SIC) system with the 1997 North American Industry Classification System (NAICS).

In this paper, we discuss our continuing research to improve the sample design methods used for our sample revisions. Specifically, we discuss decision criteria used to automate the determination of the number of strata and the take-all stratum boundaries. We also describe an improvement to our method of computing sample sizes to meet multiple reliability constraints.

2. Sample Design for BSR-2K

The BSR-2K samples were selected using a stratified random sample design, with strata determined by kind of business and size. As part of the sample design, we used two phases to compute *sample design parameters* such as the number of strata, stratum bounds, and stratum sample sizes. Stratum sample sizes were calculated using standard formulas (Cochran, 1977).

For BSR-2K, the number of strata and take-all stratum bounds were determined by an iterative process that relied on the judgment of the designer. Because this process was time consuming, we used early sampling frame data to construct the sampling frame and to design our samples. Later, we revised the sampling frame and sample designs using updated data. For future sample revisions, automating the process would reduce the time and subjectivity involved.

2.1. Creating Preliminary and Final Sampling Frames

The sampling frame for BSR-2K was comprised of records for two types of sampling units - *companies* and *Employer Identification Numbers* (*EINs*). Both companies and EINs are groups of one or more *employer establishments* under common ownership. An employer establishment is the smallest business unit at which transactions take place and payroll and employment records are kept. A company is comprised of *all* establishments under common ownership. An EIN is comprised of all establishments within a company that use the EIN to file payroll withholdings. In many cases, an EIN is identical to its parent company. For more information on the two types of sampling units, see Isaki, et al. (1976) and U.S. Census Bureau (2000).

For each trade area, sampling units were formed by aggregating sales and payroll data for employer establishments classified in the trade area. Thus, for a given company, the data from all employer establishments classified in retail were aggregated to create one or more sampling units, the data from all employer establishments classified in wholesale were aggregated to create one or more sampling units, and the data from all employer establishments classified in service were aggregated to create one or more sampling units.

Each sampling unit was assigned a *measure of size* that estimated the unit's annual sales at the time of sampling frame construction. Also, each sampling unit was assigned the kind of business, based on the most detailed NAICS industry levels for which estimates were to be published, that accounted for most of the unit's measure of size.

Using two types of sampling units is a compromise between data collection and maintenance for our samples. For data collection, the company is preferred because respondents usually have complete and up-to-date knowledge of company activity, including information on new activities. For sample maintenance, the EIN is preferred because using the EIN-based administrative data system is the most cost-effective method of identifying new entities and updating old ones. Using both company and EIN sampling units complicated our sample design work, as will be discussed in section 2.3.

Because of the time required to determine our sample design parameters, we used establishment data from the 1997 Economic Census in the first phase of our sample design work to construct the *preliminary sampling frame* and to determine initial sample design parameters, as will be described in section

---

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. We thank Douglas C. Bond, Carol V. Caldwell, and William C. Davie Jr. for their helpful comments.

2.3. Unlike prior sample revisions, the data from the Economic Census had not been fully edited. After significant corrections were made to data from the Economic Census, we revised our designs -- sometimes significantly.

In the second phase of our design work, we used establishment data from the Census Bureau's Standard Statistical Establishment List (SSEL) to create the *final sampling frame* and to determine final sample design parameters, as will be described in sections 2.3 and 2.5. The SSEL is a multi-relational database that contains a record for each employer establishment and is periodically updated with information from the Economic Census, the Company Organization Survey (COS), and administrative data. The final sampling frame was comprised of records for sampling units with business activity at the time this frame was constructed and included records for sampling units that became known to us only after the 1997 Economic Census. Because the SSEL had not been updated with fully edited data from the 1997 Economic Census or with information from the 1998 COS, we made corrections to the SSEL data and revised the final sampling frame and the final sample design parameters accordingly.

2.2. Strata Based on Design Requirements and Size

We formed *primary strata* based on the most detailed NAICS industry levels for which estimates were to be published. These primary strata were comprised of 84 retail strata, 41 wholesale strata, and 351 service strata. The retail primary strata were not used as part of the design of the sample used to produce estimates of monthly retail inventories, and the design of this sample will not be discussed in this paper. For information on NAICS, see U.S. Office of Management and Budget (1998).

Each primary stratum was then stratified by measure of size. These *measure of size substrata* were comprised of a *certainty substratum* (or take-all substratum) of companies and three to twelve *non-certainty substrata* of EINs not associated with companies in certainty substrata. While the companies in certainty substrata were to be selected with probability one, a sample of EINs was to be selected among the non-certainty substrata. Thus, a given sampling frame was comprised of both a *certainty sampling frame* of companies and a *non-certainty sampling frame* of EINs.

For each primary stratum, we determined the number of substrata, substratum bounds, and substratum sample sizes required to meet our *sample design requirements*. These sample design requirements were NAICS-based industry levels for which estimates were to be published; desired coefficients of variation (CVs) at publication levels on estimates of total sales and total wholesale inventories; approximate total sample sizes for each of the retail, wholesale, and service samples based on budget constraints; and lists of companies that were expected to have a large influence on the precision of our estimates and were to be selected with probability one.

2.3. Determining Substrata and Initial Sample Sizes

For each primary stratum, we used only data from the Economic Census in the first phase of our design work to determine the initial lower bound of the certainty substratum, the initial number of substrata, and initial substratum sample sizes. We used SSEL data in the second phase of our design work to determine final substrata and sample sizes required to meet the multiple CV constraints discussed in section 2.2.

In the first phase of our sample design work, we simultaneously determined initial sample design parameters for a given primary stratum by investigating, *independently* of all other primary strata, the effect on the initial sample size caused by varying the initial lower bound of the certainty substratum and the initial number of substrata. Because EINs associated with companies in the certainty substratum were to be excluded from the non-certainty sampling frame, the sampling frame for a given primary stratum was fixed only *after* we determined the certainty substratum. This dependence complicated our sample design work because a change in the initial lower bound of the certainty substratum caused a change in the sampling frame. As companies were added to the certainty substratum, the EINs associated with these companies were removed from the non-certainty sampling frame.

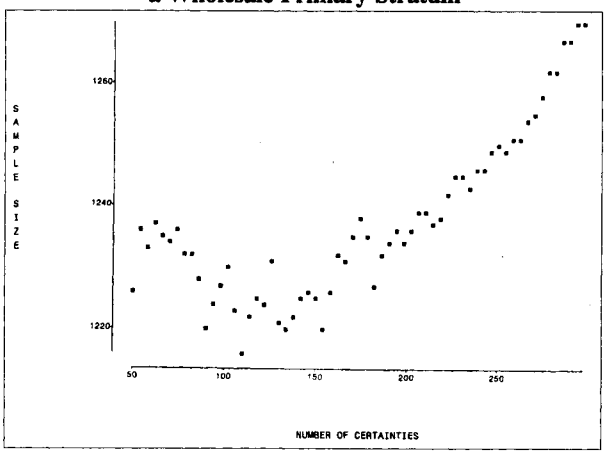
2.3.1. Determining Companies Selected with Certainty

In the first phase of our design work, we determined, for each primary stratum, the initial lower bound of the certainty substratum, or the *initial sales cutoff*, by systematically increasing the number of companies having the largest measure of size that we selected with probability one. By proceeding in this manner, for a given primary stratum, we hoped to determine the initial sales cutoff to minimize the total sample size for a fixed number of non-certainty substrata, while satisfying the desired CVs on estimates of total for the primary stratum. (For more information on how we determined substratum bounds and sample sizes, see section 2.3.2.) However, instead of finding, for each primary stratum, a distinct number of certainty companies that would minimize the sample size, we often found similar sample sizes for an interval about the number of certainty companies that gave a local minimum sample size. Such an interval was possible because of the dependence between the sampling frame and the initial sales cutoff and the imperfect relationship between the measure of size and the variables used to calculate sample sizes that will be discussed in section 2.3.2.

Figure 1 displays a graph of the sample size versus the number of certainty companies for an illustrative wholesale primary stratum. Having too few or too many certainty companies resulted in an increased sample size for the primary stratum. However, there was an interval of the number of certainty companies that gave sample sizes close to a local minimum size. In this example, having 100 to 150 certainty companies resulted in a sample size of about 1,220 units.

Figure 1.

Sample Size vs. Number of Certainty Companies for a Wholesale Primary Stratum



Because we often found similar sample sizes for an interval about the number of certainty companies that gave a local minimum sample size, we determined the initial sales cutoff to meet two goals. First, we wanted the number of certainty companies to lie in the interval to achieve an efficient design for the primary stratum. Second, for each sample, we wanted the certainty substrata to collectively contribute about the same proportion to the overall sample size as for the 1997 Business Sample Revision (BSR-97). Our analysts then identified other companies to be included in the certainty substratum because these companies were expected to have a large effect on the precision of our estimates during the life of the samples.

In the second phase of our design work, we used SSEL data to remove company records from or add company records to the certainty sampling frame. We removed the record for any company with a measure of size well below the initial sales cutoff for its primary stratum. We added the record for any company that had a measure of size greater than or equal to the initial sales cutoff for its primary stratum. We added the record for any company that had a measure of size for any of its component industries greater than or equal to the initial sales cutoff for the corresponding primary stratum. We also revised the sales cutoff for any primary stratum in which there was a large number of certainty companies that were no longer in business.

2.3.2. Determining Non-Certainty Substrata and Initial Sample Sizes

In the first phase of our design work, we used only data from the Economic Census to determine the initial number of non-certainty substrata for a given primary stratum by varying the number of these substrata. We chose a design that resulted in the smallest initial sample size required to satisfy the desired CVs on estimates of total for the primary stratum, while guaranteeing that, under Neyman allocation of the initial sample size to the substrata, no EIN from a non-certainty substratum was selected with probability one and no more than a few non-certainty substrata had sample sizes less than three units. If these conditions were not met, we changed the design to make it more efficient. Specifically, if some EINs from non-certainty substrata were selected with probability one, we lowered the initial sales cutoff. Also, if the sample sizes for more than three non-certainty substrata were less than three units, we reduced the number of non-certainty substrata from twelve in increments of three substrata until this did not occur.

We calculated non-certainty substratum bounds and initial sample sizes using the Sweet and Sigman (1995) program that allowed a modification to the Dalenius and Hodges (1959) cumulative $\sqrt{f}$ rule in which the cumulative 0.4th power of the frequency distribution of the measure of size was used to set substratum bounds. We used the cumulative 0.4th power of the frequency distribution because we found this gave slightly smaller sample sizes than the ones based on either the square root or cube root.

To compensate for using annual data to design samples that would be used to produce more variable monthly estimates and to control the influence of a given sampling unit on our estimates, we increased initial sample sizes. Controlling the influence of a given sampling unit on our estimates is important because the characteristics of a sampling unit can change during the life of the samples and affect the precision of our estimates.

We increased our initial sample sizes in two ways. First, for each non-certainty substratum, we included minimum sample size and minimum sampling rate restrictions. For each non-certainty substratum, we required the sample size to be at least three units and to result in a sampling rate greater than or equal to a minimum sampling rate that depended on the particular trade area. Second, instead of computing sample sizes for estimating total measure of size, we computed sample sizes for estimating a total that was assumed to be highly correlated with total measure of size.

For the retail, wholesale, and service samples, we computed sample sizes for estimating the total of a payroll-based regression estimate of annual sales at the time of sampling frame construction. We used a regression method (McNeil, 1977) that modeled the relationship between payroll and sales establishment data from the 1997 Economic Census at the six-digit NAICS level. For each sampling unit, we then aggregated establishment-level regression estimates that were formed by multiplying the appropriate regression coefficient by the most recent annual payroll estimate available. For the wholesale sample, we also computed sample sizes for estimating total 1997 end-of-year inventories based on data from the 1997 Economic Census. We then determined the initial sample size for a given wholesale primary stratum to be the maximum of the two sample sizes computed.

In the second phase of our design work, we used SSEL data

to determine final substratum bounds and sample sizes. Before determining these sample design parameters, we removed from the non-certainty sampling frame the records for all EINs associated with companies on the certainty sampling frame and reduced the number of substrata for primary strata in which the sample design was grossly inefficient.

2.4. Automating the Determination of Initial Designs

Recall from section 2.3 that, for BSR-2K, designer judgment was required to determine the initial sales cutoff and the initial number of substrata for each primary stratum. There are two major disadvantages associated with this method. First, a large amount of time is involved, requiring the use of early sampling frame data. Therefore, this method does not incorporate final sampling frame data and allows little flexibility in changing initial parameters on a flow basis to meet updated sample design requirements. Second, a given sample design is dependent on the judgment and experience of the person choosing the design.

Because of the disadvantages associated with the BSR-2K method of determining the initial sales cutoff and the initial number of substrata for each primary stratum, we attempted to automate the determination of these sample design parameters. To accomplish this task, we devised the following list of decision criteria based on our design experiences:

1. Include at least five companies in the certainty substratum.
2. Reduce the initial number of non-certainty substrata in increments of three substrata until no more than three substrata have sample sizes under Neyman allocation that are less than three units.
3. Add companies to the certainty substratum in increments of three until no EIN from a non-certainty substratum is selected with certainty due to Neyman allocation of the total primary stratum sample size to the substrata.
4. Add companies to the certainty substratum in increments of three as long as no more than ten sampling units are added to a local minimum sample size.
5. Require the certainty substratum to contribute at least twenty percent to the total sample size for the primary stratum.

We decided to add companies to the certainty substratum and reduce the initial number of substrata in increments of three to improve the speed of the automation.

Based on our BSR-2K sample design experiences, we wanted to limit the effect of outlier EINs on sample sizes as a starting point for automating the determination of initial sample design parameters for a given primary stratum. In the second phase of our BSR-2K design work, we discovered a relatively small number of EIN records on the final non-certainty sampling frame for which the relationship between the measure of size and the variables used to calculate sample sizes was very different from other EIN records in the respective substrata. This resulted in particular substrata having very large sample sizes, compared to the initial sample sizes computed using only data from the Economic Census. For information on the BSR-2K sample design changes due to outlier EINs, see Kinyon, et al. (2000).

To limit the effect of outlier EINs on sample sizes, we began by determining the design for each primary stratum using five companies in the certainty substratum and twelve non-certainty substrata. We removed from the non-certainty sampling frame the record for any EIN whose inclusion on the frame would have resulted in an increase of greater than one sampling unit on the total primary stratum sample size required to satisfy CV constraints at the primary-stratum level. We then added to the certainty sampling frame the company records associated with these outlier EINs and removed from the non-certainty sampling frame all EIN records associated with company records on the revised certainty sampling frame. In a future sample revision, the sampling frame data for outlier EINs would be targeted in the edit process in an effort to keep the records for some of these EINs on the non-certainty sampling frame.

In lowering the initial sales cutoff of a primary stratum by adding companies to the certainty substratum, we limited the number of EINs that had a measure of size greater than or equal to the sales cutoff and were not associated with companies in the certainty substratum. Such EINs were possible because the certainty substratum of a given primary stratum was determined *independently* of the other primary strata. Because an EIN of this type was part of a company that both was stratified in a primary stratum different from that of the EIN and was likely to be included in the certainty substratum of the company's primary stratum, we removed the records for these EINs from the non-certainty sampling frame before determining initial sample design parameters for a given primary stratum.

A problem we encountered in determining the initial sales cutoff based on the number of companies in the certainty substratum was the existence of multiple companies that had a measure of size equal to the initial sales cutoff. This was a problem because not all companies with a measure of size equal to the initial sales cutoff were added to the certainty substratum for some primary strata. Further, if such a company had only one EIN, the EIN record was dropped from the non-certainty sampling frame, as described above. To solve this problem, we allowed the number of companies in the certainty substratum of a given primary stratum to increase by increments of three companies only if there were no other companies classified in the stratum that had a measure of size equal to the initial sales cutoff. If such companies existed, we also added these companies to the certainty substratum.

Table 1 shows an example of output from the automated determination of initial sample design parameters for a retail primary stratum. Each row of the table pertains to a particular design output. The last two columns are the number of non-certainty substrata (or "EIN substrata") having sample sizes under

Neyman allocation that are less than three units and the number of non-certainty substrata in which each EIN is selected with certainty, respectively.

Table 1.
Example of Automated Determination of Initial Sample Design Parameters for a Retail Primary Stratum

Number of EIN Substrata	Number of Certainty Companies	Stratum Sample Size	Number of EIN Substrata with $n_h < 3$	Number of Certainty EIN Substrata
12	5	47	11	1
9	5	41	7	1
6	5	38	1	1
6	8	32	1	0
6	11	34	5	0
3	11	40	0	0
3	14	38	0	0
3	17	35	0	0
3	20	38	0	0
3	23	41	0	0
3	26	43	0	0

In Table 1, we begin with twelve EIN substrata and five companies in the certainty substratum for the first sample design. For the second, third, and sixth sample designs, the number of EIN substrata is reduced by three to satisfy design criterion 2. For the fourth sample design, three companies are added to the minimum number of five companies in the certainty substratum to satisfy design criterion 3. For the remaining sample designs, additional companies are added to the certainty substratum in increments of three to satisfy design criteria 4 and 5. Note that, for the penultimate sample design, the certainty substratum contributes about 56% to the total sample size and only nine units are added to the local minimum sample size of 32 by adding companies to the certainty substratum. The penultimate sample design is chosen as the final design for this primary stratum because the last sample design adds eleven units to the local minimum sample size, violating design criterion 4.

To evaluate our automated method of determining initial sample design parameters for each primary stratum, we compared the sample design parameters for the retail sample that were output from this method to those output from the BSR-2K method. For both methods, we used final sampling frame data as input and handled outlier EINs as described earlier in this section. While the automated method resulted in about 24% more certainty companies than the BSR-2K method, this method resulted in only about a 2% increase in total sample size from that resulting from the BSR-2K method. The two methods gave a different number of substrata for 28 of the 80 primary strata that were included in the test, but there were only three primary strata for which the two methods resulted in a difference of more than three substrata. Of these three primary strata, two violated design criterion 2 under the BSR-2K method.

While the automated method appears to have created similar designs to those resulting from the BSR-2K method, we should experiment with design criterion 4 because this criterion adds companies to the certainty substratum as long as the resulting sample size is within ten units of the local minimum sample size. Our goal was to have the certainty substrata contribute about the same proportion to the total retail sample size as for BSR-97. However, these substrata contributed about 34% to the total retail sample size, surpassing the targeted 25%. Also, the slightly larger total sample size associated with the automated method is most likely due to this design criterion.

2.5. Computing Final Sample Sizes to Meet Multiple Coefficient of Variation Constraints

As discussed in section 2.2, among our sample design requirements were desired CVs on estimates of total at NAICS-based publication levels. In general, these publication levels exhibited a hierarchical, or nested, structure that was based on the six-digit coding system employed by NAICS. For example, estimates were desired for the following retail industries: New Car Dealers (NAICS 441110), Automobile Dealers (NAICS 4411), Motor Vehicle and Parts Dealers (NAICS 441), and Retail Trade (NAICS 44-45).

For BSR-2K, we computed final sample sizes required to satisfy desired CVs on estimates of total at publication levels as follows. First, for each publication level, we determined the sample sizes required to satisfy the desired CVs on estimates of total using the variables described in section 2.3.2. We then calculated the corresponding sample sizes under Neyman allocation for each non-certainty substratum that contributed to the publication level. Finally, for each non-certainty substratum, we took as the final sample size the maximum of the sample sizes computed for the substratum. For the wholesale sample, sample sizes were generally larger for estimating total 1997 end-of-year inventories than for estimating total annual sales using the payroll-based estimate.

Table 2 illustrates the method used to determine final sample sizes to meet multiple CV constraints on total annual sales estimates for one of the non-certainty substrata of the New Car Dealers (NAICS 441110) retail primary stratum. In this table, the NAICS code for a given row is contained, or nested, in the NAICS codes of all rows that follow. Because the maximum sample size required for this substratum was 46.37, a sample of size 47 guaranteed that all CV constraints affecting this substratum were satisfied.

As discussed in section 2.3.2, we included minimum sample size and minimum sampling rate restrictions. For substrata whose sample sizes were *not* determined from these restrictions,

Table 2.
Determining the Required Sample Size for the Smallest Non-Certainty Substratum of NAICS 441110

NAICS Code	Desired CV (%)	Required Substratum Sample Size
441110	0.90	46.37
4411	1.79	12.58
441	1.79	11.04
44, 45	0.45	38.04

we recomputed sample sizes after removing the contribution of the restricted substrata from desired variances at publication levels. We did this because, in the substrata that we "froze" by setting their sample sizes based on the restrictions, we selected a larger sample than Neyman allocation required to meet the desired CV constraints. Therefore, smaller sample sizes were possible for substrata whose sample sizes were *not* determined from the restrictions.

A slight improvement to the method used in BSR-2K to compute final sample sizes for a given publication level is to take into account the larger sample sizes required for more detailed publication levels nested within the publication level. This improvement is an extension of the method above in which we "froze" substrata and recomputed sample sizes. To compute the sample size for a given publication level, we "freeze" the substratum sample sizes controlled by a more detailed publication-level CV constraint and recompute sample sizes for remaining substrata similar to before.

For example, suppose a given publication level is comprised of primary strata A, B, and C. Also, suppose that the sample size for stratum A is controlled by a stratum-level CV constraint, while the combined sample size for strata B and C is controlled by a publication-level CV constraint. Because the sample size for stratum A is larger than the one computed by allocating to the primary strata the sample size required to satisfy the publication-level CV constraint, a smaller combined sample size for strata B and C is possible.

A complication in extending our method of "freezing" substrata and recomputing sample sizes is that the publication-level structure for each sample is not always purely nested. Thus, it is not always clear which substrata should be "frozen" at a particular stage. For the example above, suppose there are also CV constraints at the combined strata A and B level and at the combined strata B and C level. In such an instance, the sample sizes required to satisfy the CV constraints should be examined to determine the maximum gain in efficiency resulting from "freezing" substrata and recomputing sample sizes.

We experimented with our improved method of computing final sample sizes for the retail sample, which generally has a nested publication-level structure. We computed the total sample size both with and without the improvement. Incorporating the improvement resulted in a total sample size reduction of about two percent.

3. Future Research

Research to improve the sample design methods for future sample revisions is an ongoing process. While we investigated automating the determination of the initial sample design parameters for each primary stratum and improving our method of computing final sample sizes, work remains.

Instead of using the company as a sampling unit for certainty substrata, it might be more accurate to use a different sampling unit that is more closely related to our reporting units. Because reporting units for a given company selected with certainty allow us to partition the company's data by kind of business, we should look at using kind-of-business partitions of companies when determining the certainty component for our samples.

Another area for research is the evaluation of our method of increasing sample sizes. We should examine, by kind of business, monthly variation and changes in sampling units over time to determine if the sample size increases are sufficient. This can be determined only after data is collected from respondents of our retail, wholesale, and service surveys.

4. References

Cochran, W. (1977), *Sampling Techniques*, New York: John Wiley & Sons.

Dalenius, T. and J. Hodges (1959), "Minimum Variance Stratification," *Journal of the American Statistical Association*, Vol. 54, pp. 88-101.

Isaki, C., K. Wolter, T. Sturdevant, N. Monsour, and M. Trager (1976), "Sample Redesign of the Census Bureau's Monthly Business Surveys," *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 90-98.

Kinyon, D., D. Glassbrenner, J. Black, and R. Detlefsen (2000), "Designing Business Samples Used for Surveys Conducted by the United States Bureau of the Census," paper presented at the second International Conference on Establishment Surveys, Buffalo, NY.

McNeil, D. (1977), *Interactive Data Analysis: A Practical Primer*, New York: John Wiley & Sons.

Sweet, E. and R. Sigman (1995), *User Guide for the Generalized SAS Univariate Stratification Program*, Technical Report #ESM-9504, Washington, DC: Bureau of the Census.

U.S. Census Bureau (2000), Current Business Reports, Series BR/99-A, *Annual Benchmark Report for Retail Trade: January 1990 to December 1999*, Washington, DC.

U.S. Office of Management and Budget (1998), *North American Industry Classification System: United States, 1997*, Lanham, MD: Berman Press.