# INVESTIGATION OF SELECTIVE EDITING PROCEDURES USING QUINQUENNIAL DATA[1]

Katherine Jenny Thompson and Susan Lynn Hostetter
U.S. Bureau of the Census, Department of Commerce

**Abstract**
Selective editing scores individual questionnaires based on their **potential** effects on the estimates, selecting only the cases with a high probability of impact on tabulations for analyst referral. In 2000, we conducted an investigation into the feasibility of using selective editing methods on an annual survey with the ultimate goal of developing a selective editing methodology for use in the 2002 Economic Census. This paper applies our recommended approach to quinquennial Economic Census data, used at both the macro and micro levels and concludes with recommendations for the 2002 Economic Census programs.

**Keywords:** Edit referral, score function

## 1. Introduction

Every five years, the U.S. Bureau of the Census conducts a census of businesses. This Economic Census provides an important framework and benchmark for composite measures such as the gross domestic product estimates, input/output measures, and production and price indexes. Economic Census tabulations profile the U.S. economy from the national to the local level at a more detailed industry and geographic area level than the more frequently collected statistical series that measure short-term changes in the economy. Moreover, Economic Census micro-data is used to construct frames for sample surveys and is used as input for economic modeling.

Economic Census data is reviewed in many different ways before publication. The first review is the micro-level review of edit-failing records. Typically the records are reviewed extensively at this point with thousands of records being referred to each analyst to review, one by one. After micro-review is completed, analysts begin table cell analysis (macro-review). In many programs, selected individual records undergo a third review, reconciling reported census data to data collected from the same units in current annual surveys. Finally, there is another stage of micro-review of data used for frame construction. The multiple phases of this review process can be quite fatiguing for the analysts, which in turn can affect data quality.

This study came about as an effort to improve the efficiency of the first micro- review of edit-failing records. The Economic Census is administered by nine different program areas. Currently, procedures for determining which edit-failing records should be micro-reviewed differ by program area. Our goal is to determine whether selective editing can replace these individual census programs' edit referral procedures. Selective editing determines edit referrals based on the questionnaire's **potential** effects on the estimates, isolating the edit failures with the largest expected impact by using score functions with predetermined critical values. Critical values are computed from a prior survey/census cycle so that processing is not delayed by their calculation. During the actual editing cycle, records with scores that exceed their associated critical value are referred to analysts; all others are automatically imputed. Selective editing is applied **only** to records with non-fatal edit errors (response items altered by the edit). Fatal edits – such as blank required items or

[1]The authors thank Roger Goodwin, Carol King, Richard Sigman, and Brandy Yarbrough for their helpful comments on earlier versions of this manuscript. 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.

failed industry classification – must always be resolved clerically. The term "selective editing" is a slight misnomer, or at least misleading, since all reports are machine edited. A better choice of words might be "selective referral" or "selective analyst review."

Thompson and Hostetter (2000) presents a feasibility study on using selective editing methods on an annual survey conducted by the U.S. Bureau of the Census. Quinquennial data vary more than annual data due to inflation effects, company failures, and acquisitions. Moreover, different programs have different screening criteria for analyst referrals (e.g., some programs include full-impute records in micro-review, while others do not), so the type of data used to develop critical values can greatly differ by program. Any selective editing procedure used for the Economic Census must therefore be fairly insensitive to changes over a five year period and must be robust to different referral eligibility rules. Additionally, it should not reduce the quality of the edited micro-data. This paper addresses these issues, investigating the selective editing methodology proposed in our earlier study on quinquennial data from the 1992 and 1997 Census of Construction and from the Census of Services-Sectors Businesses. We present the results of our investigation and conclude with recommendations for the 2002 Economic Census programs.

## 2. Selective Editing Methodology

Selective editing calculates a single measurement (a global score) for each respondent after fatal edit errors have been resolved. Global scores are compared to cell-specific critical values, and the cases whose global score exceeds this critical value are targeted for analyst referral. Examples of selective editing cells include industries, or size-class-categories within industry. Cases that pass all edits have a global score of zero. Critical values are computed from prior-period data using the same global score function.

Global scores combine local scores, which are calculated for individual questionnaire items. Local scores measure the magnitude of change between the reported and edited value of selected data items. Because all local scores are combined into one global score per reporting unit, local scores must be based on the same units of measurement. For example, number of employees is made comparable to other items such as annual payroll or total receipts by multiplying the employment data by an industry's average earnings ratio (wages/employment) before calculating its local score.

We used the global and local score functions recommended by Thompson and Hostetter (2000). Our local score function for each data item *i* is given by

$$LS_i = \sqrt{MAX\left(r_i \times V_i, e_i \times V_i\right)} \times z_i \tag{3.1}$$

where $r_i$ is the reported value of item *i*, $e_i$ is the edited value of item *i*, $V_i$ is the industry average for data item *i* if that item is not reported in dollars ($V_i$ = 1 otherwise), and $z_i$ is an 0/1 indicator value for reported items that are considered edit failures (definitions of edit failures vary by program). Our global score is the maximum value of the establishment's local scores multiplied by the sample weight ($\equiv$ 1 for census data). So, a large edit change to any variable can potentially cause a referral.

We used the "simulation study" approach to develop critical values (Lawrence and McKenzie, 2000). Using 1992 data, we calculate the following two statistics at five percentiles *p* (*p* = 45, 55, 65, 75, 85) of the global score distribution in the selective editing cell:

- Absolute pseudo-bias, calculated for each data item as $|\hat{Y}_C - \hat{Y}_{100}|/\hat{Y}_{100}$ , where $\hat{Y}_C$ is the estimate of the item total calculated by replacing all reported values with a global score function larger than the critical value by their edited values and leaving reported values in place for the others, and $\hat{Y}_{100}$ is the corresponding total calculated from 100% edited data (Latouche and Berthelot (1992)); and

- Referral rates, calculated as the number of establishments with a global score greater than the critical value divided by the total number of establishments in the industry.

We try to balance obtaining low pseudo-bias for each item and maintaining low referral rates (each program has its own acceptable referral rate level). When our critical values are successful, the pseudo-bias of each item at the critical value is "high" (say 0.05), the pseudo-bias at all preceding global score values is near zero, and the referral rate does not exceed the program threshold. Because analyst resources are limited, we relax our pseudo-bias criterion if necessary rather than increase our referral rate cut-off limit. We evaluate our selected 1992 data critical values by calculating corresponding pseudo-biases and referral rates on the 1997 census data.

In a production environment, the critical values must be computed before the current data is received. For the 2002 census, we would use the same methods on 1997 data to develop critical values. Our earlier paper recommended using a two-level flagging system, where records with a global score greater than the prior-period percentile *p* **must** be reviewed by an analyst, and the records whose global score is greater than the $(p\text{-}10)^{\text{th}}$ percentile (but less than the $p^{\text{th}}$ percentile) are reviewed as time permits. This approach is quite reasonable and is what we recommend using for the 1997 census, selecting *p* based on the 1992/1997 data patterns. See Section 4 for more details.

## 3. Case Study

This study uses 1992 and 1997 data from the Census of Construction Industries (CCI) and from the Census of Services-Sectors Businesses (CSSB). These two programs represent both ends of the Economic Census data product spectrum. The CCI data is collected to produce (macro-level) tabulations. CSSB data is used at both the macro and micro levels.

The CCI is a mail-out/mail back sample survey of approximately 130,000 employer businesses from the construction sector, publishing industry-level tabulations at the national and state levels. Analyst referrals were limited to records with large edit changes to non-zero reported values for at least one item [Note: automatically-corrected data items reported in the wrong units – "rounded items" – are also excluded from analyst referral]. Imputations for unreported values were not referred. We use the same referral screening criterion for our global score assignments (only eligible cases get non-zero global scores). We examined fifteen industries selected by subject-matter experts and use four highly correlated data items to develop global scores: total receipts; annual payroll; cost of construction work subcontracted out; and cost of materials, components, and supplies.

The CSSB is a mail-out/mail-back census of over four million businesses. This census comprises six trade areas: Retail Trade; Wholesale Trade; Service Industries; Transportation, Communication, and Utility Industries (Utilities); Finance, Insurance, and Real Estate (FIRE); and Auxiliary Establishments. Each trade area publishes industry-level statistics at the national, state, and county levels. Any large change from a reported value (excluding rounded items) could trigger an analyst referral, although generally analyst referrals were restricted to large establishments. In contrast to the CCI program, full-impute cases and unreported values were also candidates for analyst referrals. We examined 26 industries (six in Retail Trade; five in Wholesale Trade; seven in Services

Industries; five in FIRE; and three in Utilities) and used four data items to develop global scores: total sales/receipts, annual payroll, first quarter payroll, and number of employees. Originally, we included full-impute cases in our global score calculations. However, this greatly decreased the selective editing's effectiveness. By definition, all edit changes to full-impute cases are "large." The distributions of full-impute records are very different from the reporting/keying error distributions of reporter units. Mixing the two types of records was not worthwhile.

## 4. Industry Level Tabulation Results

The appendix presents average pseudo-biases and referral rates (and associated SEs) for 1992 and 1997 CCI and CSSB data (broken down by trade area) using the 1992 global score distribution percentiles. Pseudo-biases and referral rates exclude full-impute cases and "rounded" data items, yielding conservative estimates. For Construction, using the 65$^{th}$ percentiles of the 1992 global score distributions on 1992 data yielded average three-percent referral rates and pseudo-biases of less than one percent (all items); corresponding statistics are even lower when we applied the 1992 65$^{th}$ percentile to the 1997 data. For all of the CSSB programs except Wholesale, using the 85$^{th}$ percentiles of the 1992 global score distributions on 1992 data yielded average referral rates of four percent or less and kept most of the average pseudo-biases to less than five-percent. These patterns are repeated when the 1992 critical values are applied to the 1997 data with two exceptions: in FIRE, the average pseudo-bias for 1$^{st}$ quarter payroll is greater than 5-percent for all 1992 percentiles greater than 65; and in the Retail trade department store industry, where all 1992 percentiles yielded larger than 30-percent referral rates. For Wholesale, using the 75$^{th}$ percentile of the 1992 global score distribution seems to best balance low average pseudo-biases and referral rates on 1992 data; applying the 1992 critical value to the 1997 data improves the referral rates.

These results were suspiciously promising. We expect industry distributions of data items to change between censuses. Was it possible that distributions of reporting/keying errors could remain similar from census to census, while the actual data item distributions did not? Table 1 contains two-sample Wilcoxon test results comparing differences between 1992 and 1997 industry-level global score distributions ($\alpha$=0.05) using location shift alternatives.

Table 1: Test Results and Critical Value Comparison for Industry-Level Global Score Distributions

| Program | Total Industries | Location Shift Appropriate | Different Global Score Distributions ($\alpha$ = 0.05) | 1992 critical value percentile ($p$) less than 1997 $(p+10)^{th}$ percentile |
|---|---|---|---|---|
| CONSTRUCTION | 15 | 14 | 8 (3 with $\theta > 0$, 5 with $\theta < 0$) | 13 |
| RETAIL | 6 | 6 | 3 (3 with $\theta > 0$) | 6 |
| WHOLESALE | 5 | 5 | 4 (4 with $\theta < 0$) | 5 |
| SERVICES | 7 | 7 | 6 (4 with $\theta > 0$, 2 with $\theta < 0$) | 7 |
| FIRE | 5 | 5 | 5 (5 with $\theta < 0$) | 5 |
| UTILITIES | 3 | 2 | 1 (1 with $\theta > 0$) | 3 |

In most industries, the global score distributions are significantly different. Prior to testing, we verified that such alternatives were appropriate by viewing overlaid global score distribution function graphs and by comparing corresponding percentile differences (i.e., $p_{i,97}$ - $p_{i,92}$, where $p_i$ is a percentile. Consistently positive differences imply that the location shift $\theta$ is greater than 0 and consistently negative differences imply that $\theta < 0$). In most cases, the two distribution function graphs were indistinguishable. And, in general the percentile differences had the same sign (within industry) up to both distributions' 95$^{th}$ percentiles. When we excluded all cases above the (1992 or

1997) 95th percentile in both data sets, the Wilcoxon tests were still significant, confirming the existence of location shifts. Table 1 presents the results of these comparisons, along with location shift signs as appropriate. Notice that 16 of the 27 significantly different global score distributions have negative location shifts. However, the consistency between 1992 and 1997 referral rates and pseudo-biases within industry implies that these location shifts are generally negligible.

Previously, we concluded that global score distributions from the current and prior data collection periods had to be statistically **equivalent** for selective editing to be effective. Obviously, these results contradict this. Our original conclusion was too limited. Selective editing works well when the distributions of global scores have the same shape and a "small" location shift up to a cut-off value between consecutive time periods. The critical value for the global score should be less than this cut-off value. It is not necessary – or often even possible – for the two global score distributions to be equivalent. Selective editing predicts **where** outliers are located in the global score distribution. In each census, a high percentage of the automatically-corrected errors are keying or balancing errors (e.g., corrected transposed digits, replaced reported total with associated sum of details). These types of data corrections often have little effect on the tabulations. Consequently, it is not unreasonable to assume similar global score distributions for current period and prior period data up to a certain percentile (say, the 90th percentile).

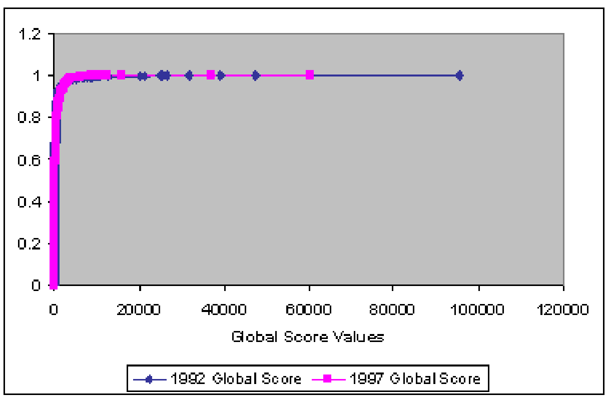


**Figure 1**: CDF Plots of 1992 and 1997 Global Score Distributions in a Construction Census Industry

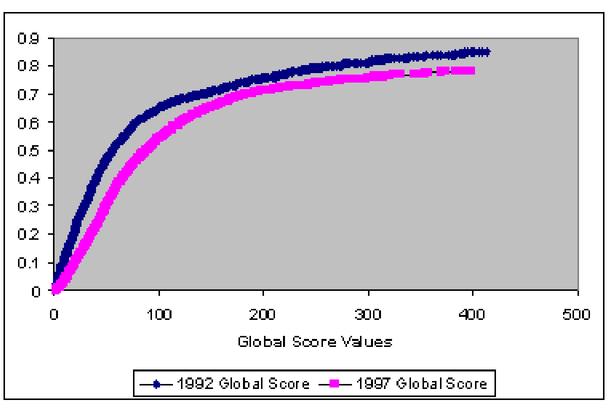


**Figure 2:** CDF Plots of Truncated 1992 and 1997 Global Score Distributions in the Same Industry

Figures 1 and 2 illustrate this, presenting the 1992 and 1997 global score distributions for a CCI industry. The two plotted distributions are statistically different. Figure 1 plots the entire cumulative density function of both score functions. It is impossible to see any differences in distribution because the two curves are so close; however, the 1992 scores are consistently lower than the 1997 scores until the 95th percentile of each distribution. Truncating each curve at a global score value of 400[2] (Figure 2) shows the positive location shift between the two curves. Because the shift is positive, no accuracy is lost using the 1992 critical value with the 1997 data. In both distributions, the true outlier observations lie beyond their respective 95th percentiles. Our critical values are set well before the 1992-data 95th percentile value to prevent overly-high pseudo-biases, so they identify all 1997 data outliers.

In our test industries, a high percentage of the significantly different global distribution functions appear to have negative location shifts (i.e., 1997 global score values tend to be smaller than the 1992 global score values). These location shifts are small; the two sets of curves are

[2]the 78th percentile of the 1997 distribution and the 85th percentile of the 1992 distribution

very close. In all but two of our industries, the 1992 critical value percentile (*p*) is consistently smaller than the $(p + 10)^{th}$ percentile of the corresponding 1997 distribution. Using a slightly lower value of *p* than warranted by the 1992 pseudo-biases/referral rates helps ensure that all of the outliers greater than the 1997 $(p + 10)^{th}$ percentile are flagged for analyst review, regardless of location shift sign. For example, to ensure that the all 2002 global score values greater than the (2002) $85^{th}$ percentile are flagged, we would flag all records with a global score greater than the 1997 $75^{th}$ percentile for definite analyst review, with all records with a global score between the 1997 $65^{th}$ and $75^{th}$ percentiles to be reviewed as time permits. If referral rates are too high (or to low), we would adjust our (1997) "must" percentile accordingly.

## 5. Investigating the Effect of Selective Editing on Micro-data

Selective editing is really designed for programs that collect data solely for tabulation purposes. The integrity of the micro-data is not a requirement. The CSSB Economic Census data are used, however, at both macro and micro levels, so it is important to examine the effect of selective editing at the micro-level as well as the macro-level. Industry-level selective editing cells are probably sufficient for accurate tabulation, but may be inadequate for frame development: for example, small establishments with "large" edit changes can greatly impact sample survey stratification and allocation. We hoped that by refining the selective editing cells, we could develop critical values that preserved both the macro and micro-data.

This section describes our investigation of this approach in the Retail, Wholesale, and Services trade areas. Census data from these three trade areas are used to construct frames for annual and monthly surveys. These surveys use company data or employer identification number (EIN) data as sample units, not individual establishments (Kinyon et al, 2000). These sample units are **aggregated** establishment data: a company is comprised of all establishments under common ownership; and an EIN sample unit is comprised of all establishments within a company that use the same EIN to file payroll withholdings. Census micro-review is **establishment-based.** Questionnaires must be edited and reviewed as received. Analysts cannot wait to receive all questionnaires from a company or an EIN to begin review. Consequently, we could not use the surveys' strata boundaries to define our smaller selective editing cells. An alternative option would have been to use the census programs' imputation cells – Services and Wholesale use legal form of organization, tax status, and type of operation to futher classify establishments within industry for imputation – but the classification variables used to define these cells were not available to us (and for Retail, the industry is the imputation cell).

Instead, we developed size-class cells (based on establishment data) within industry. Our goal was to examine the **feasibility** of this approach, not to recommend a method of developing size-class cells. In a production system, the selective editing and imputation cells should be the same to reduce parameter overhead (e.g., development, maintenance, and application). For this portion of the study, we developed a (maximum of) seven size-class-within-industry selective editing cells from the 1992 edited data. One size-class cell contained all establishments that exceeded the $90^{th}$ percentile of the industry's total annual payroll distribution, thus assuring that all "large" establishments had a chance of being reviewed. We used the Sweet and Sigman GUS program (1994) to implement the Dalenius-Hodges cum $\sqrt{f}$ rule (Cochran, 1977) with annual payroll as the stratification variable to create the other six cells. [Note: using sales to define size-classes yielded much higher and more variable industry-level pseudo-biases].

Then, we repeated the process described in section 4 to develop critical values for each cell. For simplicity, we used the same percentile (the program-specific percentile recommended in section 4) in all size-classes within an industry as a critical value. We verified our results using **industry-level** pseudo-biases and referral rates, making sure not to improve the data quality at the cost of greatly increased referral rates. Recall that we are focusing on the first stage of micro-review, which works towards production of quality industry-level tabulations given tight time-constraints. At this stage of the process, analysts confine their review to cases for which large edit changes should impact tabulations (small establishments may not receive the same scrutiny). Often, during stratification, these large establishments become self-representing. Stratification and allocation are consequently more affected by the smaller establishments. "Strata jumpers" – establishments classified into the wrong strata – can greatly affect the variances estimated from the survey data, making them much larger than the target variances used to design the sample. Similarly, "missed" large data errors can affect allocation by (falsely) increasing certain within-stratum variances.

To assess the effect of selective editing on frame construction, we constructed three separate **establishment**-level frames from our 1997 census Retail, Wholesale, and Services data. Frame 1 consisted entirely of final edited values (no selective editing), representing "truth." Frame 2 consisted of **industry-cell** selectively edited values. Edited values are used for full-impute cases, rounded cases, and cases whose global score exceeds the industry-level critical value; reported data is used otherwise (similar to the data sets used to compute absolute pseudo-biases). Finally, frame 3 consisted of **industry × size-class cell** selectively edited data, constructed similarly to the industry-cell level frame.

We compared the Neyman allocations obtained from each frame for a stratified simple random sample without replacement design. We used six strata[3] per industry, stratifying on sales/receipts (the stratification variable used for the current surveys), applying the cum $\sqrt{f}$ to define strata on the 1997 data . Given each c.v. constraint (0.01, 0.05, 0.10), we compared the sample-size produced by three different allocation variables -- sales, annual payroll, and employment – selecting the largest of the three allocations. This allowed us to assess the impact of selective editing on the micro-level for three of our four variables [Note: 1st quarter payroll is so highly correlated with annual payroll that we did not evaluate it separately].

Using selectively edited data – either industry-cell or industry by size-class cell – did not affect the strata boundaries in this test: all three sets of strata boundaries were very close. Choice of selective editing cell did, however, greatly affect allocation. Table 2 presents the average industry-level percent change in sample size from fully-edited establishment data for both the industry-cell selectively edited establishment data (frame 2) and the industry by size-class cell selectively edited establishment data (frame 3), along with the range of absolute percent difference [Note: these percent differences were all positive, except for one Services industry, whose percent change was approximately -0.08(%) for all c.v.'s]. Due to the small number of test industries, one large percent change can greatly affect a trade area's mean, so Table 2 also contains counts of industries that have less than five percent change from fully-edited data.

Notice the effect of selective editing cell on allocation. Using industry-level selective editing cells generally leads to unacceptably large increases in required sample sizes. The increased allocation is due to (relatively) large differences between selectively-edited data and fully-edited data in a **small**

[3] As recommended by Cochran (1997) for variables whose correlation is generally less than 0.95.

establishment stratum. Although these differences are proportionally small overall (hence the low pseudo-biases), they greatly increase the within-strata variability. In contrast, the allocations from industry by size-class selectively edited (frame 3) data are generally very close to those obtained from the fully-edited data. This includes Retail, whose Table 2 results are a bit misleading. The high average percent increase in all categories is caused by very poor results in one industry. If omitted, then the average percent increase for Retail ranges from 40 to 49 percent in the industry-level cells and from 0.4 to 0.5 percent in the industry by size-class cell.

Table 2: Percent Absolute Difference in Allocated Sample Sizes (Relative to Fully Edited Data)

| | c.v. = 0.01 | | c.v. = 0.05 | | c.v. = 0.10 | |
|---|---|---|---|---|---|---|
| | Industry | Industry By Size | Industry | Industry By Size | Industry | Industry By Size |
| RETAIL (6 Industries) | | | | | | |
| Mean (minimum, maximum) | 68 (3, 212) | 34 (0, 202) | 360 (5,1919) | 311 (0, 1865) | 450 (5, 2454) | 397 (0,2381) |
| No. of Industries with < 5% change | 1 | 5 | 1 | 5 | 1 | 5 |
| WHOLESALE (5 Industries) | | | | | | |
| Mean (minimum, maximum) | 21 (3, 56) | 3 (0, 9) | 63 (4, 250) | 5 (0, 13) | 72 (4, 292) | 6 (0, 14) |
| No. of Industries with < 5% change | 2 | 3 | 1 | 3 | 1 | 3 |
| SERVICES (7 Industries) | | | | | | |
| Mean (minimum, maximum) | 100 (0, 486) | 1 (0, 4) | 159 (0, 851) | 2 (0, 4) | 170 (0, 926) | 2 (0, 4) |
| No. of Industries with < 5% change | 3 | 7 | 3 | 7 | 2 | 7 |

Why were the allocation results so poor in this particular Retail industry? In **one** small-establishment stratum, there were several "large" differences between selectively-edited data and fully-edited data for annual payroll. These differences did not impact the industry-level tabulations; in fact, the industry-level pseudo-bias for the frame 3 annual payroll data is low, around three percent. So, the selective editing process here resulted in within-stratum variance approximately ten times larger than the "true" within-stratum variance (constructed from fully edited data). This problem is undetectable at the macro level. Moreover, the within-strata variability for sales and employment using the frame 3 data in the same stratum are quite reasonable. This indicates that further review of cases in small-establishment strata is probably necessary when using selective-edited microdata.

The current surveys use aggregate data, not establishment data. Small differences at the individual establishment level are potential large differences when aggregated (for example, one grocery store establishment in a chain may not comprise a large percentage of an industry tabulation, but the aggregated company data might). Table 3 examines the effect on allocation of using industry by size-class selectively edited **establishment** data to create a frame of EIN-unit data, again compared to data from an frame of EIN-units constructed from fully edited data. Once again, the selective editing had little – if any – effect on stratification.

Table 3: Percent Absolute Difference in Allocated Sample Sizes with EIN Data (Relative to Frame Constructed from Fully Edited Data)

| | c.v. = 0.01 | c.v. = 0.05 | c.v. = 0.10 |
|---|---|---|---|
| RETAIL | | | |
| Mean (minimum, maximum) | 63 (0, 376) | 333 (0, 1997) | 293 (0, 1755) |
| No. of Industries with < 5% change | 5 | 5 | 5 |
| WHOLESALE | | | |
| Mean (minimum, maximum) | 2 (0, 6) | 3 (0, 9) | 3 (0,10) |
| No. of Industries with < 5% change | 3 | 3 | 4 |
| SERVICES | | | |
| Mean (minimum, maximum) | 1 (0, 2) | 1 (0, 2) | 1 (0, 2) |
| No. of Industries with < 5% change | 7 | 7 | 7 |

The aggregated data results reinforce our earlier findings. By carefully defining our selective editing cells within industry, we reduce the number of large differences between selectively-edited and fully-edited data within size-category at the establishment level. Note that the same pattern for the problem Retail trade industry appears here.

## 6. Conclusion

This study investigated the use of our previously recommended selective editing methodology on quinquennial census data. We examined the effect of selective editing at both the macro and micro levels on data from two different programs, each of which has its own screening criteria for analyst referral cases. The results from this study confirm our earlier findings: namely, that selective editing using our recommended score function has desirable properties in most cases (low referral rates and low pseudo-biases) because of the consistent reporting/keying error patterns between censuses. The effect of inflation/deflation – a major concern – turned out to be negligible, at least in our sample industries. Moreover, our score functions and critical value selection methodology are fairly robust to different screening criteria as long as at least one data item is reported.

We strongly believe that this methodology should be pursued for the 2002 Economic Census although there are a few further areas that must be investigated for production implementation. First, our method does not work with selective-editing cells containing less than 20 observations (both in current and prior periods). This was not an issue with this study's test industries but could be in the future. Second, we need to develop size-class-within-industry selective editing cells that are consistent with imputation cells for programs that use the micro-data as well as the macro-data. After determining the selective editing cells, we can use the same approach described in section 3 to develop critical value percentiles (although implementors may not want to institute a two-tiered flagging system in the small establishment cells). Finally, we need to develop a "fall-back" plan for industries whose critical values result in overly-high or overly-low referral rates (those industries whose reporting/keying error distributions change greatly between collection period). Such fine-tuning will result in a product with wide program-application potential that will save analyst resources while preserving data quality.

## References

Cochran, W. (1977). Sampling Techniques. New York: Wiley.

Kinyon, D.; Glassbrenner, D.; Black, J.; and Detlefsen, R. (2000). Designing Business Samples for Surveys Conducted by the United States Bureau of the Census. *Proceeding of the Second Annual Conference on Establishment Surveys.*

Latouche, M. and Berthelot, J.M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics*, **8**, pp. 389-400.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. *Journal of Official Statistics*, **16**, pp. 243- 255.

Sweet, E. and Sigman, R. (1995). User Guide for the Generalized SAS Univariate Stratification Program. U.S. Bureau of the Census: ESM-9504.

Thompson, K.J. and Hostetter, S (2000). Investigation of Selective Editing Procedures for The U.S. Bureau of the Census Economic Programs. *Proceedings of the Second International Conference on Establishment Surveys.*

# Appendix: Average Pseudo-Biases and Referral Rates

| Data Items | | 45th Percentile (Pseudo bias/ standard error) | 55th Percentile (Pseudo bias/ standard error) | 65th Percentile (Pseudo bias/ standard error) | 75th Percentile (Pseudo bias/ standard error) | 85th Percentile (Pseudo bias/ standard error) |
|---|---|---|---|---|---|---|
| CONSTRUCTION | | | | | | |
| **Total Receipts** | **1992** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.03 / 0.02** |
| | **1997** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.04 / 0.01** |
| **Annual Payroll** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.03 / 0.02 | 0.45 / 0.19 |
| | 1997 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 | 0.22 / 0.07 |
| **Cost of Const. Work Out** | **1992** | **0.01 / 0.00** | **0.02 / 0.01** | **0.04 / 0.02** | **0.15 / 0.15** | **0.69 / 0.45** |
| | **1997** | **0.00 / 0.00** | **0.01 / 0.00** | **0.01 / 0.02** | **0.03 / 0.02** | **0.14 / 0.02** |
| **Cost of Mat., Comp. and Sup.** | 1992 | 0.01 / 0.00 | 0.01 / 0.00 | 0.02 / 0.00 | 0.05 / 0.00 | 0.24 / 0.06 |
| | 1997 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 | 0.02 / 0.01 | 0.09 / 0.01 |
| **Referral Rate** | **1992** | **0.04 / 0.01** | **0.03 / 0.01** | **0.03 / 0.01** | **0.02 / 0.00** | **0.01 / 0.00** |
| | **1997** | **0.06 / 0.03** | **0.05 / 0.03** | **0.04 / 0.02** | **0.03 / 0.02** | **0.02 / 0.01** |
| **RETAIL** | | | | | | |
| **Total Sales** | **1992** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.01 / 0.01** | **0.01 / 0.01** |
| | **1997** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.01 / 0.01** |
| **Annual Payroll** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.01 |
| | 1997 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 |
| **1st Quarter payroll** | **1992** | **0.01 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** | **0.02 / 0.02** |
| | **1997** | **0.02 / 0.01** | **0.02 / 0.01** | **0.02 / 0.01** | **0.02 / 0.01** | **0.02 / 0.01** |
| **Total Employment** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |
| | 1997 | 0.01 / 0.00 | 0.01 / 0.00 | 0.02 / 0.00 | 0.02 / 0.01 | 0.02 / 0.00 |
| **Referral Rate** | **1992** | **0.05 / 0.01** | **0.04 / 0.01** | **0.03 / 0.00** | **0.02 / 0.00** | **0.01 / 0.00** |
| | **1997** | **0.14 / 0.01** | **0.12 / 0.01** | **0.11 / 0.01** | **0.10 / 0.01** | **0.08 / 0.01** |
| **WHOLESALE** | | | | | | |
| **Total Sales** | **1992** | **0.00 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** | **0.02 / 0.01** | **0.03 / 0.01** |
| | **1997** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** |
| **Annual Payroll** | 1992 | 0.00 / 0.00 | 0.01 / 0.01 | 0.01 / 0.01 | 0.01 / 0.01 | 0.03 / 0.01 |
| | 1997 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 |
| **1st Quarter payroll** | **1992** | **0.01 / 0.00** | **0.02 / 0.01** | **0.02 / 0.01** | **0.02 / 0.01** | **0.06 / 0.02** |
| | **1997** | **0.03 / 0.02** | **0.04 / 0.02** | **0.04 / 0.02** | **0.04 / 0.02** | **0.04 / 0.02** |
| **Total Employment** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 |
| | 1997 | 0.00 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 |
| **Referral Rate** | **1992** | **0.10 / 0.04** | **0.08 / 0.03** | **0.06 / 0.02** | **0.04 / 0.02** | **0.03 / 0.01** |
| | **1997** | **0.07 / 0.00** | **0.06 / 0.00** | **0.05 / 0.01** | **0.03 / 0.01** | **0.02 / 0.01** |
| **SERVICES** | | | | | | |
| **Total Sales** | **1992** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.02 / 0.00** | **0.04 / 0.00** |
| | **1997** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.01 / 0.01** | **0.01 / 0.01** |
| **Annual Payroll** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |
| | 1997 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 |
| **1st Quarter payroll** | **1992** | **0.01 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** |
| | **1997** | **0.02 / 0.01** | **0.02 / 0.01** | **0.02 / 0.01** | **0.02 / 0.01** | 0.02 / 0.00 |
| **Total Employment** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |
| | 1997 | 0.00 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 |
| **Referral Rate** | **1992** | **0.04 / 0.03** | **0.04 / 0.02** | **0.03 / 0.02** | **0.02 / 0.01** | **0.01 / 0.01** |
| | **1997** | **0.07 / 0.04** | **0.06 / 0.03** | **0.05 / 0.03** | **0.04 / 0.02** | **0.03 / 0.02** |
| **FINANCE, INSURANCE,REAL ESTATE** | | | | | | |
| **Total Sales** | **1992** | **0.00 / 0.00** | **0.00 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** |
| | **1997** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** |
| **Annual Payroll** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 |
| | 1997 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.02 / 0.01 |
| **1st Quarter payroll** | **1992** | **0.01 / 0.01** | **0.01 / 0.01** | **0.01 / 0.00** | **0.01 / 0.00** | **0.01 / 0.00** |
| | **1997** | **0.04 / 0.01** | **0.04 / 0.01** | **0.05 / 0.01** | **0.06 / 0.01** | **0.06 / 0.01** |
| **Total Employment** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 | 0.01 / 0.00 |
| | 1997 | 0.01 / 0.00 | 0.01 / 0.00 | 0.02 / 0.01 | 0.02 / 0.01 | 0.03 / 0.01 |
| **Referral Rate** | **1992** | **0.07 / 0.01** | **0.06 / 0.00** | **0.04 / 0.00** | **0.03 / 0.00** | **0.02 / 0.00** |
| | **1997** | **0.11 / 0.01** | **0.09 / 0.01** | **0.06 / 0.01** | **0.05 / 0.01** | **0.02 / 0.01** |
| **UTILITIES** | | | | | | |
| **Total Sales** | **1992** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.01 / 0.01** | **0.03 / 0.04** |
| | **1997** | **0.01 / 0.00** | **0.01 / 0.00** | **0.02 / 0.01** | **0.02 / 0.02** | **0.04 / 0.05** |
| **Annual Payroll** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.01 | 0.01 / 0.01 | 0.02 / 0.02 |
| | 1997 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.01 | 0.00 / 0.00 | 0.02 / 0.02 |
| **1st Quarter payroll** | **1992** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.00 / 0.00** | **0.01 / 0.01** |
| | **1997** | **0.01 / 0.01** | **0.01 / 0.01** | **0.02 / 0.02** | **0.02 / 0.02** | **0.04 / 0.03** |
| **Total Employment** | 1992 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.01 | 0.01 / 0.01 |
| | 1997 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.01 | 0.01 / 0.01 | 0.03 / 0.04 |
| **Referral Rate** | **1992** | **0.14 / 0.10** | **0.11 / 0.08** | **0.09 / 0.07** | **0.06 / 0.05** | **0.04 / 0.03** |
| | **1997** | **0.15 / 0.06** | **0.12 / 0.05** | **0.11 / 0.05** | **0.09 / 0.04** | **0.05 / 0.02** |