Journal of Official Statistics, Vol. 15, No. 4, 1999, pp. 517–535

# Statistical Methods for Developing Ratio Edit Tolerances for Economic Data

*Katherine Jenny Thompson and Richard S. Sigman*[1]

The U.S. Census Bureau developed general-purpose ratio edit software for use by the ten sectors of the 1997 Economic Census. This software requires explicit bounds (tolerances) for each ratio edit. We investigated statistical methods of automatically setting tolerance limits, examining three methods: robust estimation (15% trimmed mean and standard deviation); resistant fences (EDA method based on first and third quartiles and interquartile range); and gap analysis (Distance Measurement Algorithm for the Selection of Outliers, D_MASO). We also developed an approach for symmetrizing skewed distributions of ratios using power transformations prior to tolerance development. We evaluated these methods on two sets of historical data: the 1994 Annual Survey of Manufactures (ASM) and the 1992 Business Census. In both data sets, we achieved success with some variation of resistant fences and recommend that this methodology be used in the absence of subject-matter expertise or known mathematical bounds on a ratio relationship.

*Key words:* EDA (Exploratory Data Analysis); resistant; robust; ratio edit.

## 1. Introduction

Key data items collected by the economic programs of the U.S. Bureau of the Census are subjected to ratio edits as a part of the overall data-review process. In a ratio edit, the ratio of two highly correlated items is compared to upper and lower bounds, known as tolerances. Ratios outside the tolerances are edit failures, and one or both of the items in an edit-failing ratio are either imputed or flagged for analyst review. The effectiveness of the ratio edit is therefore dependent on the tolerances.

From a subject-matter analyst's perspective, ratio edits are appealing because it is difficult to evaluate the ‘‘reasonableness’’ of a data item's value by itself. By comparing an item to other related values in the questionnaire, the analyst can determine if the response appears valid. For example, the ratio of total annual hours to total employees should be approximately 2,000 (40 hours a work week multiplied by 50 work weeks a year). Ratio edit systems are equally appealing from a mathematical perspective. By augmenting explicitly defined ratio edits with implied ratio edits, a record containing edit failures can be corrected by applying a set covering procedure to the set of edit-failing

[1] U.S. Bureau of the Census, Economic Statistical Methods and Programming Division, Washington, DC 20233, U.S.A. E-mail: katherine.j.thompson@ccmail.census.gov and richard.s.sigman@ccmail.census.gov

**Acknowledgments:**We thank Jock Black, William C. Davie, Jr., Lisa Draper, Vicki Farri, Nash Monsour, two anonymous referees, and an associate editor for their useful comments. This article reports the research and analysis undertaken by staff at the U.S. Bureau of the Census. It has undergone a more limited review than official U.S. Census Bureau publications and is released to inform interested parties of research and to encourage discussion.

ratios (Fellegi and Holt 1976; Greenberg 1986). This procedure determines the minimal number of edit-failing data items that must be imputed to obtain a consistent record.

In the past year, ratio editing has become increasingly important at the U.S. Census Bureau. A team of analysts, computer scientists, and statisticians has developed a general-purpose ratio edit module based on the U.S. Census Bureau's original SPEER edit system (Structured Programs for Economic Editing and Referrals) for use by all ten economic census sectors in 1997. The SPEER system, which utilizes the Fellegi-Holt model of editing continuous data described above, has been used successfully at the U.S. Census Bureau in various forms since 1984 (Greenberg, Draper, and Petkunas 1990; Winkler and Draper 1997). Many analysts increased their use of ratio edits because of the new software. Thus, not only did old tolerances need to be updated or regenerated, but entirely new sets of tolerances had to be developed. Because some censuses use ''live'' (new) data for their tolerance development, often the parameter generation had a tight time constraint as well (a two-week turnaround for some censuses).

Traditionally, the most difficult part of ratio editing has been developing reasonable bounds. Overly tight bounds lead to a high proportion of false edit rejects, in turn resulting in over-imputation or fruitless analyst review. Overly wide bounds retain erroneous data, which can affect the final tabulations. Occasionally, bounds on a ratio can be determined by a mathematical rule: for example, the ratio of total annual hours to total employees cannot be less than zero or more than 8,784 (24 hours a day multiplied by 366 days in a leap year). More often, there is no set of known boundaries on ratio relationships. Instead, the ratio edit can be viewed as a no-intercept regression model, where the numerator is the dependent variable. In these cases, statistical methods should be used to develop tolerances.

When historical data are available, the development of statistical ratio edit tolerances begins with data analysis. Ratio edit tolerances separate a distribution of ratios into two regions: an acceptance region and an outlier region. Determining ratio edit tolerances via data analysis thus falls into the category of univariate outlier detection. In discussing outlier detection, it is important to distinguish between legitimate outliers, i.e., unusual values that are accurately recorded, and genuine errors, i.e., values that are inaccurately recorded and yield suspicious-looking results. For us, ''good'' items are those that are accurately recorded (however extreme) and ''bad'' items are those that are inaccurately recorded.

We conducted research to find a statistical method of developing ratio edit tolerances that works well for different sets of economic data. Our statistical objective was to find a technique that balanced the goals of maximizing the number of rejected bad items and minimizing the number of rejected good items in our data sets (using the historical edit outcome to classify the ratios). Our operational objective was to find a method that was easy to implement and flexible. We compared three different tolerance development methodologies, exploring variations within each to see how capable the procedure was of adapting to the nuances of different data sets. Two of the methodologies considered, robust methods and gap analysis, had been used with varying success at the U.S. Census Bureau for other economic reports. The other methodology, resistant fences, was suggested by Exploratory Data Analysis (EDA) literature.

Most of the techniques that we considered assume the ratios are symmetrically distributed. Distributions of ratios created from skewed data may or may not be symmetric, and so in Section 2 we present our approach for symmetrizing skewed distributions of ratios prior to setting the tolerances. In Section 3, we describe the outlier detection methods considered. Section 4 describes our evaluation of these methods. Section 5 discusses these results and presents our recommendation. Section 6 contains final remarks.

## 2. Symmetrizing Skewed Distributions for Tolerance Development

### *2.1. Background*

In general, distributions of economic data are positively skewed. However, most of the statistical methods we considered for tolerance development implicitly assume symmetric distributions. Often, approximate symmetry can be achieved using power transformations on the original data. A power transformation with parameter equal to $p$ is a function of the following form:

$$T_p(x) = \begin{cases} x^p & (p > 0) \\ \log(x) & (p = 0) \\ -x^p & (p < 0) \end{cases} \tag{1}$$

The negative sign before the expression for $p < 0$ guarantees that $T_p(x)$ has the same limit approaching from both left and right as $p \rightarrow 0$.

For long-tailed distributions, power transformations such as the natural logarithm ($p = 0$) or the square root ($p = 0.5$) are useful, since they expand the lower data values and shrink the spread of larger data values. The logarithm transformation has another appealing property: absolute differences on the log scale correspond to percentage differences on the original scale. However, the logarithm transformation cannot be used when the distributions contain legitimate zero values.

To estimate $p$, we first employed a modification of an EDA method for determining $p$ described by Hoaglin, Mosteller, and Tukey (1983), then applied the natural logarithm transformation to the same distribution of ratios, and selected the transformation that produced the smallest absolute value of the sample skewness coefficient (including a comparison to the skewness coefficient of the original distribution). We included the comparison to the natural logarithm transformation because the EDA method's resistance breaks down when the data set contains more than the expected number of outliers (see Section 2.2). As would be expected, we did not symmetrize every distribution of ratios. Several of the distributions contain legitimate outliers and remain skewed even after being transformed. Moreover, a symmetrizing power transform does not exist for some distributions.

Initial sets of ratio edit tolerances are obtained from the transformed distributions. The inverse power transformation is applied to the initial limits to obtain the tolerances actually used in the edit system. The inverse-transformed tolerances will not be symmetric on the original scale.

### *2.2. EDA method for symmetrizing skewed distributions*

Most of the skewed distributions we examined could be easily symmetrized by applying the natural logarithm transformation to the ratios, i.e., by applying (1) with $p = 0$. To evaluate other power transformations, we obtained $p$ with the modification of Hoaglin et al.'s (1983) transformation plot for symmetry described below.

Given an ordered sample of size $n$, let $M$ be the median of the sample, and $x_L$ and $x_U$ represent the lower and upper values of a set $i$ of $k$-percent approximate quantiles. The EDA transformation plot for symmetry places

$$x_{vi} = ((x_U + x_L)/2) - M \tag{2}$$

on the vertical axis ($v$), and

$$x_{hi} = ((x_U - M)^2 + (M - x_L)^2)/(4M) \tag{3}$$

on the horizontal axis ($h$), so that

$$\text{slope}_i = x_{vi}/x_{hi} \tag{4}$$

If the resultant graph is nearly linear, that is, slope $\approx c$, then $p = 1 - c$. We determined a value for $p$ without graphing the points, by setting $p =$ 1-median$(x_{vi}/x_{hi})$. Although Hoaglin recommends rounding $p$ to the nearest multiple of $\frac{1}{2}$, we did not employ this rounding.

The EDA procedure resistance breaks down (exceeds the breakdown point) when the actual number of outliers exceeds the number of observations in one (or both) of the quantiles with the smallest tail probability. The breakdown point can be raised by decreasing the number of quantiles employed. However, the lower the breakdown point, the more data points available to calculate the median slope.

To minimize the effect of multiple outliers, we deleted extreme observations from the untransformed sample before estimating the slope from the quantiles using the resistant outer fences rules described in Section 3. We refer to the reduced sample size as $n$. To maximize the resistance, we linked sample size with expected number of outliers before breakdown by using the sets of quantiles specified in Table 1.

Our variation of the EDA method allows for four expected outliers in the subsetted data set before breakdown.

## 3. Outlier Detection Methods

We examined three approaches to setting tolerance limits: a robust approach (fifteen-percent trimmed mean and standard deviation); an outlier-resistant approach (resistant fences); and a gap analysis approach (Distance Measurement Algorithm for the Selection of Outliers, (D_MASO)). These methods are described below.

*Table 1. Sets of quantiles for EDA transformation plot for symmetry*

| Range of $n$ | Breakdown point | Quantiles used |
|---|---|---|
| 33–64 | 1/8 | 1/4, 1/8 |
| 65–128 | 1/16 | 1/4, 1/8, 1/16 |
| 129–256 | 1/32 | 1/4, 1/8, 1/16, 1/32 |
| 257–512 | 1/64 | 1/4, 1/8, 1/16, 1/32, 1/64 |
| 513–∞ | 1/128 | 1/4, 1/8, 1/16, 1/32, 1/64, 1/128 |

### 3.1. Robust mean and standard deviation

In an earlier economic census editing cycle, the first step of tolerance development was to eliminate all of the ratios in a class that were more than two standard deviations from the mean and to base the tolerances on the resultant data set. We attempted an analogous technique using robust estimates of the population mean and standard deviation. Other case studies have employed robust estimation techniques to develop tolerances (Mazur 1989; Pierce and Gillis 1995).

We used the fifteen-percent trimmed mean[2] as our robust estimate of the population mean and a robust estimate of population standard deviation based on the Winsorized sum of squared deviations (a consistent estimator of the variance of the trimmed mean, as shown in Gross (1976)). A more conservative approach such as the five-percent trimmed mean would have yielded wider tolerances (hence fewer false edit rejects) but would have been less robust.

If the distribution of ratios (or transformed ratios) is approximately symmetric, the robust estimates of the population mean and standard deviation define a robust confidence interval for the mean (Gross 1976; Mazur 1989). Based on the principle that almost any distribution is ‘‘normal in the middle,’’ the interval $(\bar{x}_k \pm 2\sigma_k)$ should include roughly ninety-five percent and $(\bar{x}_k \pm 3\sigma_k)$ should include roughly ninety-nine percent of the observations. The analyst thus has some control over the number of edit rejects. Using $(\bar{x}_k \pm 2\sigma_k)$ to define the acceptable region amounts to a more liberal rule, while using $(\bar{x}_k \pm 3\sigma_k)$ amounts to a more conservative rule in terms of the number of cases flagged for review.

### 3.2. Resistant fences

Resistant methods are insensitive to outlying observations in the distribution (Hoaglin et al. 1983), producing results that change only slightly when a part of the data is replaced by new (entirely different) numbers. West (1995) proposed using an EDA outlier detection method called resistant fences to develop tolerances. Resistant fences rules are based on sample quartiles. Given an ordered distribution of ratios, let $q_{25} =$ the first quartile, $q_{75} =$ the third quartile, and $H = q_{75} - q_{25}$, the interquartile range. The resistant fences rules define outliers as ratios less than $q_{25} - k^*H$ or greater than $q_{75} + k^*H$, where $k$ is a constant. The inner fences rule sets $k$ equal to 1.5 and is the rule used to define the standard Tukey boxplot. The outer fences rule sets $k$ equal to 3. A compromise rule – the middle fences rule – sets $k$ equal to 2. We calculated the quartiles by setting the cumulative probability level for the $i$th order statistic $x_{(i)}$ equal to $i/(n+1)$ as recommended by Hoaglin and Iglewicz (1987).

The resistant fences rules implicitly assume – but do not require – that the ratios are symmetrically distributed. We applied the resistant fences rules to both transformed and untransformed data.

### 3.3. Distance measurement algorithm for selection of outliers (D_MASO)

The D_MASO gap analysis approach was developed at the U.S. Census Bureau to develop

[2] That is, we eliminated the $k$ smallest and largest observations from our distributions, where $k$ is fifteen percent of the sample size $n$.

ratio edit tolerances for the 1992 Enterprise report (Oh, Paletz, Kim, and Salyers 1994). D_MASO examines successive ratios of ordered explicit ratios, considering proportional distances between adjacent ordered observations to find potential edit bounds. An ''unusually'' large gap between adjacent observations at either end of the distribution may indicate that the observations between the gap and the end of the distribution are outliers. The observation on the ''center side'' of the gap is a potential tolerance. The user specifies the maximum percentage of observations that can be labeled as outliers and specifies a cut-off value for distance comparisons. Appendix A outlines the D_MASO procedure.

## 4. Determining Ratio Edit Tolerances

### *4.1. The ratio edit as a hypothesis test*

A ratio edit is a hypothesis test, in which the null hypothesis is that both data fields in the ratio are correct (Pierce and Gillis 1995). One rejects the null hypothesis when the ratio falls outside of the tolerances. Given this definition, we can define Type I and Type II error for each ratio edit. A Type I error flags a ratio value as an error when it is in fact correct: these are good ratios that fall outside of the tolerances. A Type II error flags a ratio as correct when it is in fact an error: these are bad ratios that lie inside of the tolerances. Type I error increases unnecessary analyst work and is generally controlled by widening the tolerances. However, the wider the tolerances, the greater the probability of Type II error.

Some caution must be used in defining Type II error. Only a portion of the Type II error for an individual ratio test is controlled by the tolerance limits. With ratio edits, there is usually an ''inlier'' set of bad ratios, where an inlier is defined as a bad ratio whose value is consistent with the rest of the distribution. For example, ratio edits rarely identify rounding errors: if both items are reported in the wrong units (e.g., thousands instead of units), the ratio value will be acceptable even though both data items are scaled incorrectly. Furthermore, an item in a ratio is often involved in more than one ratio test. An item value that is acceptable in one ratio relationship edit may be unacceptable in another.

Because item values are often tested in more than one ratio edit, the individual ratio edit Type II error is a poor measure of the overall proportion of uncorrected (unidentified) bad items left remaining in the edited data. Consequently, we define the Type II error of an entire set of ratio edits as bad data that passes all containing ratio edits. At the ratio-edit level, the all-ratio-test Type II error for the complete set of ratio edits is the number of bad ratios that are not outside of any tolerances. At the individual data item level, the all-item Type II error for the complete set of ratio edits is the bad data items that are not contained in any ratios outside the tolerances, i.e., the bad items that are not identified as outliers by any of the ratio edit tests.

We can also examine the power of a set of ratio edits. For outlier detection, the power is the probability of correctly concluding that a bad ratio is an outlier. The power of a set of ratio edits is the proportion of bad ratios that fall outside of one or more tolerances.

An alternative measure for evaluating the tolerances for a given ratio test is the hit rate – the ratio of the number of bad ratios outside of the tolerances to the total number of ratios outside of the tolerances (Granquist 1995). This is an important measure in evaluating the

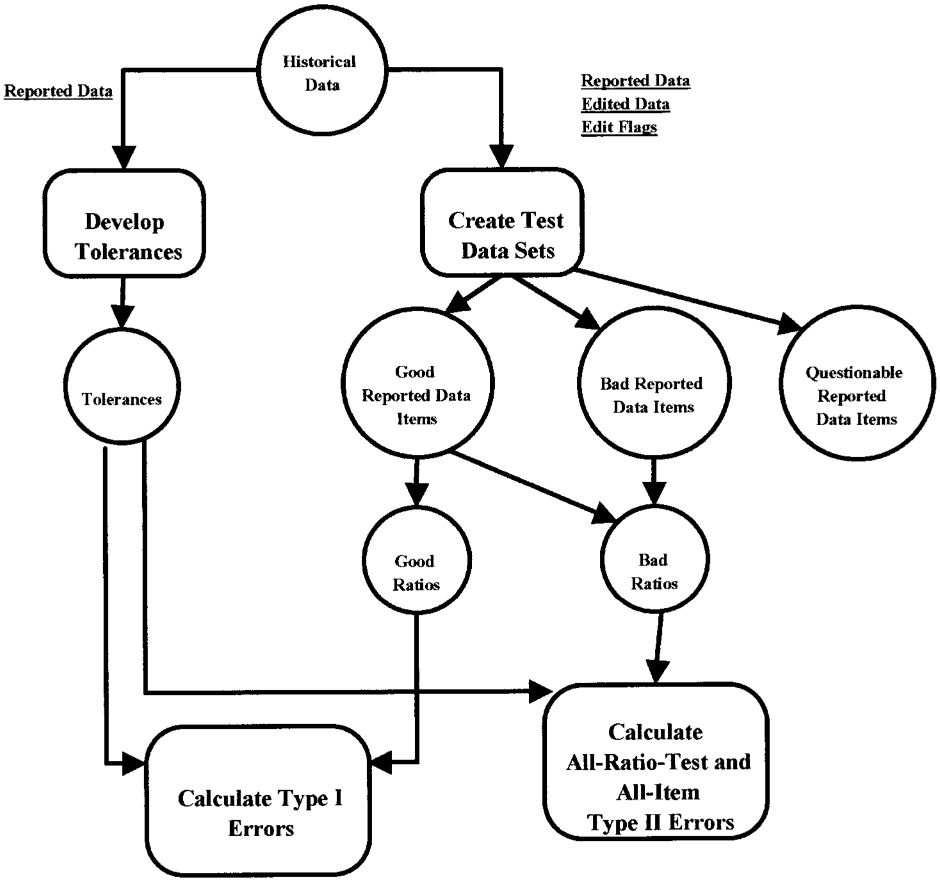

*Fig. 1. Classification and use of historical data*

operational effectiveness of an edit. The higher the hit rate, the more erroneous observations correctly flagged.

To examine the different methods, we used historical data (see Section 4.2.) and classified each ratio based on its historical edit outcomes.[3] First, we considered each nonblank/zero data item separately, using the reported data item, the edited data item, and the data item's edit action flag to classify each data item as good, bad, or questionable. Good data items had legitimate reported values, bad data items had unacceptable reported values, and questionable data items had reported values which were impossible to classify as good or bad. For example, in certain cases, the reported and edited value are not equal, and the item is flagged as having been imputed and then analyst corrected. On average, roughly five percent of the nonzero data items were flagged as questionable.

A ratio is good if both the numerator and the denominator are flagged as good. A ratio is bad if either the numerator or the denominator is flagged as bad. Ratios that contain blank or questionable values were excluded from all evaluations. The flowchart in Figure 1 shows how we classified and used the historical data.

[3] Data items may have been subjected to other edits besides ratio edits, such as balance edits and consistency edits.

We assume that the edited historical data are correct. This assumption is not unreasonable. Edited economic census data are subjected to a separate outlier detection program, giving extreme values two chances of review. Additionally, large establishments are often flagged for analyst review regardless of edit outcome. And, in many cases, aggregate census cell totals are compared to corresponding administrative data totals. It is of course possible that undetected erroneous variables are included in the edited data (and classified as ''good'' in our system), especially among small establishments.

### 4.2. *Historical data sources*

We examined two sets of historical data: the 1994 Annual Survey of Manufactures (ASM) and the 1992 Business Census. Our data included only full-year reported establishments and excluded all fully imputed cases.

The ASM is a mail-out/mail-back survey representing all establishments that received a form in the previous census of manufactures. The survey provides detailed annual statistics on the location, activities, and products of approximately 58,000 U.S. manufactures. Prior to ratio-editing, the ASM reported data undergoes some clerical edits. We used this partially edited data to develop the tolerances. Twelve of the ASM ratio edits require statistically developed tolerances and the same set of ratio edits is used in each standard industrial classification (SIC).

The Business Census is a quinquennial mail-out/mail-back census that covers five trade areas: Retail Trade; Wholesale Trade; Service Industries; Transportation, Communication, and Utility Industries (Utilities); and Finance, Insurance, and Real Estate (FIRE). Data are collected on approximately 150 different questionnaires, and over four million census forms are mailed out. The Business Census data are ratio edited without a prior clerical edit. Administrative data is substituted for blank data whenever possible to develop tolerances, so we used reported and administrative data to develop the tolerances. Some trade areas classify the establishments within SIC by legal form of organization, type of operation, and tax status. We used the trade area classifications for our evaluation, but refer to each classification as an SIC. Each trade area in the Business Census employs a common set of core ratio edits. Four of these ratio edits require statistically developed tolerances. We performed our evaluation by trade area within census for the four statistically determined ratio edits.

### 4.3. *Evaluation methodology*

We generated nine sets of edit tolerances per ratio test in each SIC: two sets of robust edit limits for symmetrized distributions; three sets of resistant edit limits (inner, middle, and outer fences limits) for symmetrized and unsymmetrized distributions (six sets total); and one set of gap analysis (D_MASO) edit limits. We used sixteen SICs for the ASM evaluation and thirty SICs for the Business Census evaluation.

Our objective was to find a technique that balanced the goals of maximizing the number of rejected bad items and minimizing the number of rejected good items. Appendix B presents the average Type I error rates for each ratio test, the all-ratio-test Type II error rate, and the all-item Type II error rate obtained using each tolerance development method on the 1994 ASM data. Appendix C presents the same statistics for the 1992 Business Census data by trade area. The Type I error rate for each ratio test within an SIC is the number of good ratios outside of the tolerances divided by the total number of good ratios. The all-ratio-test Type II error rate is the number of bad ratios inside the tolerances of all containing tests divided by the total number of bad ratios. The all-item Type II error rate is the number of bad items which are not outside of any tolerance limits divided by the total number of bad items for the trade area (Business Census) or survey (ASM). The all-item Type II error rate is always larger than the all-ratio-test Type II error rate.

The Type I error rates should be interpreted cautiously. Usually there were very few good ratios outside of the tolerances, and a difference of one or two cases gives the misleading appearance of a large change in the error rate.

### 4.3.1. Comparison of robust and resistant methods: symmetrized data

To eliminate the method that performed the worst overall in terms of Type I error (false reject rate), we first examined the robust and resistant methods on symmetrized distributions, considering a high proportion of ratio tests with a Type I error larger than 0.10 unacceptable. Across the board, the tolerances generated with two robust standard deviations were too narrow. The robust methods performed poorly because the tails of the symmetrized distributions were heavier than those of a normal distribution with $\sigma^2$ equal to that estimated by the Winsorized variance estimator, so $\pm 2\sigma$ did not cover the expected 95 percent. Moreover, in the ASM data set, the $\pm 3\sigma$ did not cover the expected 99 percent.

The appeal of the robust methods was the potential for control over Type I error. This was not the case with our data sets. Furthermore, the tolerances generated with three robust standard deviations were similar to those generated with the resistant middle fences ($k = 2$), so there was no apparent advantage in further pursuing the robust estimation techniques. Mazur (1989) reached the same conclusion with livestock slaughter data.

### 4.3.2. Comparison of resistant fences methods: symmetrized and unsymmetrized data

We next considered the resistant methods separately on symmetrized and unsymmetrized distributions. For each SIC/ratio, we selected a ''best'' resistant fence rule for the unsymmetrized data and for the symmetrized data. We then examined whether symmetrizing was necessary for the historical data used. For more details, see Thompson and Sigman (1996).

For most of the ASM ratios, the same resistant fence rule (outer fences) worked best on both the symmetrized and the unsymmetrized data. In fact, the tolerances generated from both data sets were similar. Although the ASM distributions of ratios are generally positively skewed, the degree of skewness is often not severe[4] as in some economic applications. The symmetrizing compressed the ASM distributions of ratios but did not dramatically change their shape. Often, the skewness of the transformed distributions was not substantially reduced because the longer tail consisted entirely of legitimate outliers. Consequently, the tolerances developed from the symmetrized data were only slightly narrower than those from the unsymmetrized data.

In general, the tolerances calculated from the unsymmetrized ASM data yielded tests with slightly higher hit rates (proportion of rejected ratios that were bad) than those calculated from symmetrized data. Consequently, the Type I error rate (proportion of rejected good ratios) is also lower for the unsymmetrized distributions on a case by

[4] The median skewness coefficients from each distribution of ratios (within SIC) ranged from 1.16 to 8.02.

case basis. Moreover, the all-ratio Type II error rates were essentially the same for the symmetrized and unsymmetrized distributions.

In contrast with the ASM results, the resistant methods perform quite differently on the symmetrized and unsymmetrized distributions of Business Census data. The Business Census data is highly positively skewed[5] more so than the ASM data. Because of the degree of skewness, the interquartile range ($H$) is generally larger than $(q_{25} - x_{(1)})$ for the unsymmetrized data: the lower bound is almost always negative, and the upper bound is near the center of the distribution. For most of our data sets, applying the natural logarithm transformation to the distributions of ratios corrected the skewness. When the resistant rules were applied to the symmetrized data, the tolerances were generally near the ends of the distributions. Thus, the symmetrizing decreased the Type I error rate and yielded approximately the same all-ratio Type II error rate.

Applying the resistant outer fences ($k = 3$) to the original unsymmetrized distribution of ratios usually worked well for the ASM data. This was rarely the case with the Business Census data. Although symmetrizing the distributions improved the hit rate of the tests, a distance of three interquartile ranges from the upper and lower quartiles was not often sufficient. In three of the five trade areas, the Type I error rate was too high for $k = 3$ (often larger than 0.05). We found that specifying four interquartile ranges ($k = 4$) improved the Type I error rate with very little loss in individual hit rates or total power for the census of Retail Trade, the census of Service Industries, and the census of Transportation, Communication, and Utilities Industries.

### 4.3.3. Comparison of resistant fences and D_MASO procedure

The D_MASO algorithm was developed at the U.S. Census Bureau to generate tolerances for the 1992 Enterprise Report (Oh et al. 1994). There are some key differences between the D_MASO approach and the resistant approaches. First, the D_MASO procedure does not use a probability model. Second, the D_MASO procedure looks for separate groups of observations to determine outlier zones, rather than looking for extreme observations. Finally, the user specifies *a priori* the maximum proportion of the data that can be labeled as outliers by the D_MASO procedure.

For our application, we specified a maximum outlier proportion of five percent per tail and used the default cut-off factors of 1.2 for the lower and upper tail as specified in Oh et al. (1994), finding the algorithm fairly insensitive to the cut-off percent when the default cut-off factors were between 1.2 and 3. In fact, the selection of the cut-off factor had a larger effect on the tolerance limits than the cut-off percent. We compared the D_MASO procedure to the most successful resistant procedure for each historical data set: outer fences with unsymmetrized ASM data; and outer fences ($k = 3$) or ''big'' fences ($k = 4$) with symmetrized Business Census data, depending on trade area. See Thompson and Sigman (1996) for more details.

For most of the ASM ratios, the resistant fences usually performed better than D_MASO. In the few cases where the D_MASO bounds were clearly superior, the original distributions are very positively skewed; the resistant fences bounds were too narrow and

[5] Median skewness coefficients for each distribution of ratios ranged from 14.61 to 38.94. Broken down by trade area, the median skewness ranged from 14.61 to 20.53 (Wholesale); from 20.14 to 33.20 (Retail); from 24.93 to 38.94 (Services); from 18.08 to 27.88 (Utilities); and from 19.06 to 23.77 (FIRE).

were negative at the lower end. In general, however, the ASM resistant fences tolerances were usually slightly wider than the D_MASO tolerances and identified the same bad ratios. Consequently, the Type I error rate is usually higher for D_MASO. The power is about the same for the two methods, although the hit rate is generally higher for the resistant fences tolerances. However, the difference in error rates and hit rates between the two methods is usually caused by a small number (two or three) of rejected good ratios.

For the Business Census data, the resistant methods outperformed D_MASO in three trade areas: Wholesale, Utilities, and FIRE. In the other two trade areas, the resistant methods and D_MASO tied in terms of overall performance. The D_MASO procedure limits the number of observations that can be flagged as outliers, so the procedure begins at the tail ends of the distribution. In cases where the interquartile range was small and the range of the distribution of ratios was large, the resistant bounds were much narrower than the D_MASO bounds. In these cases, the D_MASO bounds outperformed the resistant fences bounds by a large margin in terms of rejected good ratios (Type I error).

## 5. Discussion

We examined variations of three different approaches to setting tolerance limits. None of these approaches incorporated specialized subject-matter knowledge of the distribution of ratios. Analysts who work with economic data develop an expert understanding of the distributions of ratios in a given industry. A statistical methodology cannot replace this knowledge. However, it can serve as a good starting point, especially when there is no known mathematical relationship to rely upon.

Outlier detection methods can fail to work properly when more than one outlier is present. Problems that arise in the presence of multiple outliers are of two types: masking and swamping. Masking occurs when the presence of several outliers makes each individual outlier difficult to detect. Swamping occurs when multiple outliers cause the procedure to erroneously flag too many observations as outliers. These two problems can adversely impact tolerance development.

The resistant fences rules were designed to reduce masking. Because they are based on quartiles, they have a breakdown point of approximately 25%. Swamping must be controlled by the choice of $k$, the number of interquartile ranges between the quartiles and the fences. However, Hidiroglou and Berthelot (1986) note that resistant fences methods are not free from masking.

They cite two specific masking effects, both of which were present in our analysis. First, if the distribution is very positively skewed, then outliers on the left tail of the distribution are undetectable (as they are with generated negative lower tolerances for data that is always non-negative). Second, the resistant fences method does not make a specific provision for the size of the establishment, and the variability of ratios for small establishments is larger than the variability of ratios for large establishments. If the establishment size varies widely within an SIC, then too many small units will be flagged as outliers, and not enough large units will be considered. Hidiroglou and Berthelot refer to this as the ''size masking effect.''

Hidiroglou and Berthelot address these two problems with their statistical edit. This procedure transforms a distribution of ratios using a nonlinear symmetrizing transformation based on the median ratio, then multiplies the transformed observations by the larger value in the individual ratio raised to a power $U$, where $0 \leq U \leq 1$ ($U$ provides a control of the importance associated with the magnitude of the data). Quantiles are calculated from the resultant products (called effects), and the ratio edit tolerances are derived from these quantiles. Each ratio edit is performed separately. Items whose effects lie outside the tolerances are flagged for review/imputation.

The Hidiroglou-Berthelot statistical edit does not, however, easily adapt to an editing system consisting of multiple ratio edits which must be satisfied simultaneously, such as the one employed at the U.S. Census Bureau. The effects from each set of ratios are nonlinear functions of the original data and are scaled differently for different ratios, so finding an imputation solution that satisfies all edits is difficult and may not even be possible. Fortunately, the resistant fences techniques can be modified to address these two masking problems. If a highly positively skewed distribution of ratios has a heavy tail, then symmetrize the distribution with a power transformation and determine the initial tolerances with respect to the transformed data. The inverse-transformed final tolerances should locate outliers in both tails of the distribution [recall that the left and right tolerances are different distances from the median]. The ‘‘size masking effect’’ can be controlled by subgrouping establishments within cells (industries) by an establishment-size categorical variable.

D_MASO was designed to reduce swamping. The user specifies the maximum percentage of the data set that can be identified as outliers. In this case, the masking is controlled by the choice of lower and upper cut-off factors. The larger gaps in proportional distances are usually due to the smaller establishments, so the D_MASO algorithm is also prone to the size masking effect.

In terms of outlier detection, the resistant methods were the most consistently successful in balancing minimum Type I error and maximum power, working best on approximately symmetric distributions. After fine-tuning some of the values of $k$, we were able to develop tolerances with low Type I error rates and reasonable power for most of the ASM and the Business Census ratio edits. As always, there is a trade-off between Type I error and Type II error: by minimizing the Type I error rate, we increase the Type II error rate for the set of ratio edits and correspondingly reduce the power of the set of ratio tests.

D_MASO worked quite differently for the two sets of historical data. Usually, the D_MASO bounds were too tight with the ASM data: the algorithm appeared to be quite prone to swamping. This result surprised us because it was counter to the design of the algorithm. We expected the D_MASO bounds to be wider than the resistant bounds in most cases. There was no clear pattern for the Business Census data.

The appeal of the D_MASO approach is the user’s control over the maximum number of outliers. From a statistical perspective, this is not necessarily a strength. Deciding *a priori* on the number of outliers that can be detected has quality implications for the final edited data; tabulations may use data that contains several unexamined erroneous observations. If the number of establishments is large, then the Type II error rate can have a significant effect on the final tabulations.

For us, D_MASO was a ‘‘black box.’’ We did not have an intuitive understanding of the ordered distribution of gaps for a ratio and had a difficult time relating the D_MASO breaks to histograms of ratios in an SIC. In contrast, the resistant fences rules were fairly intuitive. This approach takes the shape of the center of the distribution into consideration, without making parametric assumptions about the tails. From an operational perspective, the resistant fences methods are flawed because they do not allow explicit control over the number of flagged outliers; if $k = 1$, up to fifty percent of the sample can be flagged as outliers. In practice, however, we found that the percent of observations in the rejection region could be controlled through the choice of $k$.

There are some theoretical results for normally distributed data that reinforce our conclusion. Hoaglin and Iglewicz (1987) provide details of the expected some-outside rate (fraction of the samples that will flag at least one observation as outlying) and the outside rate per observation (fraction of observations that are flagged as outliers by chance): the outer fences rule had an outside rate per observation of less that 0.5% for $n < 10$, and a some-outside rate of 1% for $n < 20$. Research by Hoaglin and Iglewicz (1987) on fixing the some-outside rate per sample found that for normally-distributed data, with sample sizes of $n > 300$, the some-outside rate was approximately ten percent for $k = 2.2$ and approximately five percent for $k = 2.4$.

Our success with the resistant outer fences has since been validated on other data sets: in particular, the 1997 Business Census (Graham 1998), the 1997 Census of Construction Industries (Kornbau 1997), and the September 1997 Hog Report (Todaro 1998). Provided that the distribution under consideration is unimodal and has a nonzero interquartile range ($H$), the resistant outer fences rules yield a good set of initial tolerances.

The necessity of symmetrizing is debatable. Symmetrizing distributions of ratios can be time-consuming and computer intensive. The effort is justified only when it reduces the Type I error rates and increases the power. However, if the same – or better – results can be obtained without first symmetrizing the data, then the additional effort is not justified. In the case of the Business Census data, the additional effort was worthwhile, but not in other data sets. And across the board, we have found that symmetrizing very small distributions (e.g., less than thirty observations) results in unusably wide tolerances. In practice, we recommend examining the degree of skewness of a representative set of distributions of ratios and examining the composition of the longer tail in a positively skewed distribution before considering power transformations.

Moreover, there are alternative versions of resistant fences rules for asymmetric distributions. Lanska and Kryscio (1997) propose using the distance between the first quartile ($q_{25}$) and the median for the lower fence and the distance between the third quartile ($q_{75}$) and the median for the upper fence instead of the interquartile range. This elongates the fences in the direction of the skewness of the distribution. While we have not investigated this method on our data sets, we believe that it is worth future consideration.

## 6. Conclusions

In this article, we have examined a variety of methods for developing ratio edit tolerances. Based on the results of our evaluation, we recommend using EDA resistant fences procedures to develop an initial set of ratio edit tolerances. If several of the distributions of ratios being edited are highly positively skewed with heavy tails, then consider combining the resistant fences techniques with the symmetrizing procedure described in Section 2 to obtain an initial set of tolerances, followed by inverse transforming the initial bounds to obtain final bounds.

Developers of ratio edit tolerances have some control over the development methodology that we propose. They can examine the distributions of the ratios and modify the development approach until it works well on the data at hand. They can work to reduce the size masking effect by subgrouping establishments within cell (industry) by size category. Implementation – even with symmetrizing – is easy and fast. Thus, this methodology appeals to statisticians and analysts alike. As an example, the 1992 Business Census used a maximum of fifteen ratio edits in 1992 (Wholesale Trade); the other Business Censuses used six ratio edits. For the 1997 Business Census, they used 65 ratios for Wholesale, 26 for Retail, 57 for Services, 64 for FIRE, and 79 for Utilities (two new censuses – Auxiliary Establishments and Outlying Areas – will also use ratio edits). This increase would not be possible without a mechanism for developing reasonable tolerances quickly (the Business Census develops ''warm deck'' parameters from live data during the Census).

No matter how much we refine a statistical methodology, however, statistical methods cannot automatically provide the ''best'' tolerances in every case. Statistical methods cannot replace subject matter expertise and common sense. As a general rule, a mathematical relationship that governs the upper and lower bounds of a ratio edit should preempt any statistical techniques. For example, the Business Census tests the ratio of Annual Payroll to First Quarter Payroll. Logically, the lower bound of this ratio is one. When we used the resistant fences methods to generate tolerances, our lower tolerances were as low as 0.85. A ratio value of 0.86 would not be flagged as an outlier for this distribution. However, one of the two items being edited is obviously wrong.

Developing good ratio edit tolerances is an iterative process. Our proposed approach provides an initial set of parameters. Data users should examine these parameters in conjunction with the microdata and modify them accordingly.

## Appendix A

*Steps in the D_MASO procedure*

Given a class variable and an explicit ratio:

1) Sort the usable observations (ratios) in ascending order.
2) Calculate ratios of the ordered ratios (proportional ratios).
3) Calculate upper and lower cut-off values. The cut-off value is the median of the proportional ratios multiplied by user-specified cut-off factors.
4) Flag tolerance at the lower end of the distribution. Starting at observation 1, compare the proportional ratio to the lower cut-off value. If the proportional value exceeds the cut-off, flag the next observation as a candidate bound. The lower tolerance is the innermost flagged observation (the flagged value closest to the center). If nothing is flagged, then the default lower tolerance is the first ratio.
5) Flag tolerance at upper end of the distribution. Starting at observation $n$, compare the proportional ratio to the upper cut-off value. If the proportional value exceeds the cut-off, flag observation $(n - 1)$ as a candidate bound. The upper tolerance is the innermost flagged observation (the flagged value closest to the center). If nothing is flagged, then the default upper tolerance is the last ratio.

## Appendix B

Average Type I Error Rates
1994 Annual Survey of Manufacturers

| RATIO EDIT | Symmetrized | | | | | Unsymmetrized | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Robust | | Resistant | | | Resistant | | | Gap |
| | Two SDs | Three SDs | Inner Fences | Middle Fences | Outer Fences | Inner Fences | Middle Fences | Outer Fences | D_MASO Default |
| cm/vs | .076 | .023 | .020 | .009 | .005 | .017 | .007 | .004 | .025 |
| le/sw | .061 | .018 | .026 | .010 | .002 | .026 | .019 | .006 | .024 |
| ow/oe | .086 | .030 | .024 | .006 | .002 | .029 | .012 | .007 | .036 |
| ph/pw | .079 | .035 | .062 | .042 | .014 | .045 | .025 | .010 | .014 |
| sw/te | .075 | .018 | .016 | .007 | .004 | .019 | .008 | .004 | .012 |
| sw/vs | .091 | .031 | .033 | .018 | .008 | .036 | .021 | .011 | .038 |
| tib/vs | .079 | .020 | .015 | .005 | .001 | .057 | .042 | .021 | .036 |
| tie/tib | .139 | .065 | .078 | .047 | .019 | .081 | .051 | .028 | .042 |
| tie/vs | .075 | .013 | .010 | .005 | .002 | .058 | .041 | .026 | .037 |
| vp/sw | .074 | .016 | .020 | .011 | .001 | .026 | .015 | .005 | .038 |
| ww/ph | .062 | .027 | .031 | .021 | .011 | .029 | .020 | .014 | .014 |
| ww/pw | .079 | .022 | .019 | .006 | .001 | .019 | .010 | .006 | .021 |
| Average | .081 | .027 | .030 | .015 | .006 | .037 | .023 | .012 | .028 |

Type II Error Rates
1994 Annual Survey of Manufacturers

| RATIO EDIT | Symmetrized | | | | | Unsymmetrized | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Robust | | Resistant | | | Resistant | | | Gap |
| | Two SDs | Three SDs | Inner Fences | Middle Fences | Outer Fences | Inner Fences | Middle Fences | Outer Fences | D_MASO Default |
| All-Ratio-Test | .544 | .710 | .694 | .777 | .858 | .715 | .773 | .850 | .804 |
| All-Item | .589 | .746 | .733 | .816 | .891 | .768 | .828 | .907 | .834 |

| *Mnemonic* | *Description* | *Mnemonic* | *Description* |
|---|---|---|---|
| TE | Total Employment | LE | Legally Required Supplemental Labor Costs |
| PW | Production Workers | VP | Voluntary Supplemental Labor Costs |
| OE | Other Employees | PH | Total Plant Hours |
| SW | Total Salaries and Wages | CM | Total Cost of Materials |
| WW | Production Workers' Wages | TIB | Beginning Total Inventories |
| OW | Other Workers' Wages | TIE | Ending Total Inventories |
| VS | Total Value of Products Shipped | | |

## Appendix C

Average Type I Error Rates
1992 Business Census

| | | Symmetrized | | | | | Unsymmetrized | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Robust | | Resistant | | | Resistant | | | Gap |
| | RATIO EDIT | Two SDs | Three SDs | Inner Fences | Middle Fences | Outer Fences | Inner Fences | Middle Fences | Outer Fences | D_MASO Default |
| WHOLESALE | APR/EMP | .100 | .034 | .027 | .012 | .002 | .044 | .023 | .012 | .010 |
| | QPR/EMP | .112 | .042 | .043 | .018 | .002 | .046 | .032 | .016 | .010 |
| | SLS/EMP | .101 | .035 | .033 | .023 | .005 | .075 | .056 | .042 | .008 |
| | SLS/QPR | .100 | .023 | .022 | .018 | .014 | .082 | .064 | .048 | .009 |
| | AVERAGE | .103 | .033 | .031 | .018 | .006 | .062 | .044 | .030 | .009 |
| RETAIL | APR/EMP | .081 | .025 | .027 | .012 | .004 | .035 | .022 | .010 | .005 |
| | QPR/EMP | .070 | .020 | .021 | .007 | .002 | .025 | .015 | .006 | .005 |
| | SLS/EMP | .095 | .027 | .028 | .013 | .002 | .050 | .033 | .018 | .009 |
| | SLS/QPR | .119 | .050 | .054 | .032 | .012 | .079 | .058 | .036 | .004 |
| | AVERAGE | .091 | .030 | .033 | .016 | .005 | .047 | .032 | .018 | .006 |
| SERVICES | APR/EMP | .089 | .027 | .027 | .010 | .002 | .032 | .017 | .007 | .010 |
| | QPR/EMP | .082 | .022 | .023 | .007 | .001 | .033 | .016 | .006 | .007 |
| | SLS/EMP | .087 | .022 | .023 | .009 | .001 | .055 | .039 | .022 | .003 |
| | SLS/QPR | .077 | .023 | .023 | .013 | .005 | .072 | .054 | .035 | .002 |
| | AVERAGE | .084 | .024 | .024 | .010 | .002 | .048 | .032 | .018 | .006 |
| UTILITIES | APR/EMP | .092 | .029 | .030 | .013 | .002 | .046 | .029 | .012 | .009 |
| | QPR/EMP | .073 | .017 | .017 | .007 | .001 | .033 | .019 | .009 | .006 |
| | SLS/EMP | .089 | .019 | .021 | .008 | .001 | .077 | .054 | .032 | .010 |
| | SLS/QPR | .092 | .027 | .028 | .013 | .002 | .087 | .067 | .042 | .004 |
| | AVERAGE | .087 | .023 | .024 | .010 | .002 | .061 | .042 | .024 | .007 |
| FIRE | APR/EMP | .109 | .048 | .044 | .020 | .005 | .038 | .026 | .016 | .011 |
| | QPR/EMP | .097 | .033 | .045 | .016 | .007 | .040 | .026 | .016 | .008 |
| | SLS/EMP | .074 | .010 | .019 | .003 | .000 | .081 | .065 | .034 | .014 |
| | SLS/QPR | .104 | .022 | .024 | .010 | .003 | .080 | .070 | .046 | .018 |
| | AVERAGE | .096 | .028 | .033 | .013 | .004 | .060 | .047 | .028 | .013 |

## Appendix C (continued)

Type II Error Rates
1992 Business Census

| | | Symmetrized | | | | | Unsymmetrized | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Robust | | Resistant | | | Resistant | | | Gap |
| | RATIO EDIT | Two SDs | Three SDs | Inner Fences | Middle Fences | Outer Fences | Inner Fences | Middle Fences | Outer Fences | D_MASO Default |
| WHOLESALE | ALL-RATIO TEST | .516 | .643 | .643 | .701 | .794 | .754 | .783 | .817 | .727 |
| | ALL-ITEM | .580 | .688 | .685 | .729 | .802 | .778 | .808 | .843 | .749 |
| RETAIL | ALL-RATIO TEST | .382 | .579 | .560 | .645 | .751 | .723 | .749 | .787 | .794 |
| | ALL-ITEM | .491 | .631 | .618 | .687 | .780 | .814 | .827 | .852 | .815 |
| SERVICES | ALL-RATIO TEST | .446 | .609 | .602 | .694 | .818 | .799 | .828 | .863 | .890 |
| | ALL-ITEM | .522 | .671 | .666 | .745 | .853 | .862 | .884 | .906 | .905 |
| UTILITIES | ALL-RATIO TEST | .340 | .596 | .576 | .736 | .862 | .747 | .774 | .803 | .848 |
| | ALL-ITEM | .461 | .707 | .698 | .823 | .900 | .808 | .834 | .855 | .880 |
| FIRE | ALL-RATIO | .367 | .522 | .506 | .588 | .690 | .653 | .671 | .689 | .704 |
| | ALL-ITEM | .477 | .602 | .587 | .665 | .758 | .733 | .749 | .757 | .757 |

| *Mnemonic* | *Description* |
|---|---|
| SLS | Total Sales |
| EMP | Total Employment |
| APR | Annual Payroll |
| QPR | First Quarter Payroll |

## 7. References

Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. New York: John Wiley and Sons.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Editing and Imputation. Journal of the American Statistical Association, 71, 17–35.

Graham, R. (1998). Developing Ratio Edit Parameters with SAS® for the Economic Census. Proceedings of the Northeast and Southeast SAS Users Group.

Granquist, L. (1995). Improving the Traditional Editing Process. In Business Survey Methods, eds. Cox et al., Chapter 21, 385–401, New York: John Wiley and Sons.

Greenberg, B. (1986). The Use of Implied Edits and Set Covering in Automated Data Editing. Technical Report Census/SRD/RR-86/02, Washington, DC: U.S. Bureau of the Census.

Greenberg, B., Draper, L., and Petkunas, T. (1990). On-Line Capabilities of SPEER. Proceedings of the Statistics Canada Symposium, Statistics Canada, 235–243.

Gross, A.M. (1976). Confidence Interval Robustness with Long-Tailed Symmetric Distributions. Journal of the American Statistical Association, 71, 409–419.

Hidiroglou, M.A. and Berthelot, J.M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12, 73–83.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (eds.) (1983). Understanding Robust and Exploratory Data Analysis. New York: Wiley.

Hoaglin, D.C. and Iglewicz, B. (1987). Fine-Tuning Some Resistant Rules for Outlier Labeling. Journal of the American Statistical Organization, 83, 1147–1149.

Hoaglin, D.C., Iglewicz, B., and Tukey, J.W. (1986). Performance of Some Resistant Rules for Outlier Labeling. Journal of the American Statistical Organization, 81, 991–999.

Kornbau, Michael (1997). 1997 Census of Construction Industries: Edit and Imputation Parameters. Unpublished internal memorandum, U.S. Bureau of the Census, Washington, D.C.

Lanska, D.J. and Kryscio, R.J. (1997). Modified Box Plots for Asymmetric Distributions. Poster Session at the Joint Meetings of the American Statistical Association.

Mazur, C. (1989). A Statistical Edit for Livestock Slaughter Data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 221–226.

Oh, S., Paletz, D., Kim, J.J-I., and Salyers, E. (1994). Development of Edit Parameters for 1992 Economic Census Enterprise Reports. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1144–1149.

Pierce, D.A. and Gillis, L.B. (1995). Time Series and Cross Section Edits with Applications to Federal Reserve Deposit Reports. Statistical Policy Working Paper 23, Part 1 of 3. Washington, DC: Office of Management and Budget.

Thompson, K.J. and Sigman, R.S. (1996). Evaluation of Statistical Methods for Developing Ratio Edit Module Parameters. Technical report #ESM-9610. Washington, DC: U.S. Bureau of the Census.

Todaro, T.A. (1998). Adapting the SPEER Edit System to Edit Hog Data in the National Agricultural Statistics Service's Quarterly Agricultural Surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 570–575.

U.S. Bureau of the Census (1993). Economic Data Programs. Washington, DC: Internal documentation.

West, S.A. (1995). Discussion. Statistical Policy Working Paper 23, Part 1 of 3. Washington, DC: Office of Management and Budget.

Winkler, W. and Draper, L. (1997). The 'SPEER' Edit System. Proceedings of the Conference of European Statisticians, Section on Statistical Data Editing (Methods and Techniques, Volume 2), United Nations Statistical Commission and Economic Commission for Europe, 51–55.

Received February 1997
Revised January 1999