Statistical Methods Used To Detect Cell-level and Respondent-level Outliers in the 2002 Economic Census of the Services Sector

Richard S. Sigman
U.S. Census Bureau

Keywords: Resistant fence, Historical ratio, Influential data

1. Background

The United States Census Bureau conducts an economic census every five years in years ending in two or seven. The primary reporting unit for the Census Bureau's Economic Census is an individual business *establishment*, which is formally defined as an economic unit, generally at a single physical location, where business is conducted or where services or industrial operations are performed. Examples include a mine, factory, warehouse, sales office, grocery store, bank, hotel, movie theatre, doctor's office, museum, and central administrative office. In industries that do not maintain financial information for individual establishments, however, the Census Bureau has defined *Alternative Reporting Units (ARU's)*, which are consolidated units made up of two or more establishments. In mining, utilities, insurance, finance, and some information industries, companies report detailed financial information for their ARU's while a separate form collects employment and payroll information for each establishment. (Mesenbourg, et. al. 2003)

The Census Bureau uses the North American Industry Classification System (NAICS) to classify businesses that participate in the Economic Census. The NAICS assigns a six-digit number to a business's U.S. industry, a five-digit number to a NAICS industry, and a four-digit number to an industry group. Three-digit NAICS codes identify sub-sectors of sectors of the economy, and two digit NAICS codes identify sectors of the economy.

The Census Bureau summarizes Economic Census data by computing totals for summary cells defined by NAICS codes and geography and then releases summary data to the public on a flow basis. The first release of summarized data is the Advance Report, containing national-level sector totals. This is followed by the Industry Series reports, which contain national-level industry totals, and the Geographic Area Series (GAS) reports, which contain industry totals for states, counties, and those places that have a 2000 Decennial Census population of 2,500 or more. Consisting of 882 individual state-by-sector reports, the GAS reports summarize approximately 1,200 NAICS codes for more than 10,000 geographic identifiers (51 states + 3,141 counties + 6,920 places). Though there are more than 12 million possible cells in the 882 GAS reports, the number of published cells is only about 2.5 million. One reason the number of published cells is less than the number if possible cells is that for many of the possible combinations of state, county, place and NAICS classification there are no associated reporting units. Another reason is that the Census Bureau does not publish summary data for cells in which it may be possible for someone to determine the data reported by an individual business from the summary data. The number of data items, which we will refer to as *cell values*, summarized for each GAS cell varies from four to eight, with the exact number depending on the economic sector.

Prior to the release of an Economic Census publication, subject-matter experts use a number of data editing techniques to detect and resolve any potential reporting or processing errors that may be present in Economic Census data. This paper focuses on the application of one of these techniques—called *outlier-cell analysis*—to the 2002 Economic Census GAS reports. Outlier-cell analysis is an example of *macroediting*, which is a technique for detecting errors in the data for individual reporting units through the analysis of aggregated data (Granquist, 1991).

2. Outlier Analysis in Prior Economic Censuses

Because the Census Bureau publishes the 882 GAS reports on a flow basis, the amount of time available to review and make needed data corrections prior to the

[1] The work described in this report was performed while the author was an employee of the U.S. Census Bureau. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

release of a GAS report is often only a few days. This requires that the pre-publication review procedures, such as outlier-cell analysis, be capable of quickly reviewing many thousands of cells and cell values. The outlier-cell analysis procedure consists of the following three steps:

- Detection of outlier cells
- Identification of influential reporting units
- Determination if reporting or processing errors are present

The first step of outlier-cell analysis labels as *outlier cells* those cells that have extreme *historic cell ratios* or extreme *current cell ratios*. A historic cell ratio is the ratio of the cell total for a particular data item calculated from current-census data to the cell total for the same cell and same data item calculated from prior-census data. A current cell ratio, on the other hand, is the ratio of a total for one data item to the cell total for a different data item in which both totals are calculated from current-census data. During the planning phase of each Economic Census, subject-matter experts decide on the numerator and denominator variables based on their knowledge of expected economic relationships between the items collected for each sector.

The second step of outlier-cell analysis focuses on the reporting units within outlier cells. This step identifies the high *influence* reporting units among those associated with each outlier cell. If θ is computed from data from the set of reporting units $\{u_j : j \in A\}$, then the influence of unit j^* on θ, denoted $I_\theta(j^*)$, is $I_\theta(j^*) = \theta - \theta_{(j^*)}$, where $\theta_{(j^*)}$ is calculated from $\{u_j : j \in A, j \neq j^*\}$. Define

C = the set of reporting units contained in a given summary cell in the current census

P = the set of reporting units contained in the same summary cell in the prior census

$$B = C \cup P$$

$$x_{C,j}^{(i)} = \begin{cases} \text{current data for item i and unit j} & j \in C \\ 0 & j \notin C \end{cases}$$

$$x_{P,j}^{(i)} = \begin{cases} \text{prior data for item i and unit j} & j \in P \\ 0 & j \notin P \end{cases}$$

$$T_U^{(i)} = \sum_{j \in B} x_{U,j}^{(i)}$$

$$T_{U(j^*)}^{(i)} = \sum_{\substack{j \in B \\ j \neq j^*}} x_{U,j}^{(i)} = T_U^{(i)} - x_{U,j^*}^{(i)}$$

for $U=C$ and $U=P$. Then

$$R_{hist}^{(i)} = \text{historic ratio for item } i = T_C^{(i)} / T_P^{(i)}$$

and

$$R_{curr}^{(i,i')} = \text{current ratio between items } i \text{ and } i' \\ = T_C^{(i)} / T_C^{(i')} \ .$$

For historic ratios,

$$\begin{aligned} I_{hist}^{(i)}(j^*) &= \text{influence of unit } j^* \\ &= T_C^{(i)} / T_P^{(i)} - T_{C(j^*)}^{(i)} / T_{P(j^*)}^{(i)} \\ &= [T_C^{Ii)} - T_P^{(i)}] / T_P^{(i)} \\ &\quad - [T_{C(j^*)}^{(i)} - T_{P(j^*)}^{(i)}] / [T_P^{(i)} - (T_P^{(i)} - T_{P(j^*)})] \end{aligned}$$

Because are interested in the influence of a reporting unit's current-census data, we ignore the term $T_P^{(i)} - T_{P(j^*)}^{(i)}$ as it represents prior-census influence. Hence,

$$\begin{aligned} I_{hist}^{(i)}(j^*) &= \text{current - census influence of unit } j^* \\ &= (x_{C,j^*}^{(i)} - x_{P,j^*}^{(i)}) / T_P^{(i)} \qquad (1) \end{aligned}$$

$$\sum_{j \in B} I_{hist}^{(i)} = (T_C^{(i)} - T_P^{(i)}) / T_P^{(i)} = R_{hist}^{(i)} \text{ -1 } .$$

Hence, the influence of a reporting unit's current-census data on the historical ratio is equal to the reporting unit's contribution to the relative change between the prior-census total and the current-census total. Consequently, if $R_{hist} > 1$, the high influence reporting units are those with large (positive) value of I_{hist}; and if $R_{hist} < 1$, the high influence reporting units are those with the smallest (negative) values of $I_{hist.}$ Moreover, it is easy to show that the influence of a group of reporting units on a historical ratio is the sum over the group of the influences of the individual reporting units. For current ratios,

$$\begin{aligned} I_{curr}^{(i,i')}(j^*) &= \text{influence of reporting unit } j^* \text{ on } R_{curr}^{(i,i')} \\ &= T_C^{(i)} / T_C^{(i')} - T_{C(j^*)}^{(i)} / T_{C(j^*)}^{(i')} . \end{aligned}$$

It can be shown that if $0 \leq x_{C,j^*} << T_C$, then

$$I_{curr}^{(i,i')} \approx (x_{C,j^*}^{(i)} - R_{curr}^{(i,i')} x_{C,j^*}^{(i')}) / T_C^{(i')} \qquad (2)$$

(Proof available from the author.) Note that

$$\sum_{j \in C} I_{curr}^{(i,i')} \approx 0 \ .$$

The GAS cell values are non-negative. From equation (2) it follows that reporting units that have high influence in causing a current cell ratio to be large will have large numerator cell values and/or small denominator cell values; whereas reporting units that have high influence in causing a current cell ratio to be small will have small numerator cell values and/or large denominator cell values. As is the case for the historical ratio, it is easy to show that the influence of a group of reporting units on a current ratio is the sum over the group of the influences of the individual reporting units.

In the third step of outlier-cell analysis, subject-matter experts determine if there are any reporting or processing errors associated with the high influence reporting units in outlier cells. If it is determined such errors are present, the subject matter experts make necessary corrections to the data.

2.1 1992 Economic Census

Braam (1992) and Shoemaker (1993) describe how the 1992 Economic Census identified outlier cells by first calculating two score values for each historic cell ratio and one score value for each current cell ratio. Let *k* index NAICS codes, and let $g\in G$ index geographical identifiers, where for state-level cells *G* is the set of all states in the U.S., for county-level cells *G* is the of all the counties in a particular state, and for place-level cells *G* is the set of all places in a particular state. The following average historic ratio across NAICS codes was calculated for each $g\in G$

$$R_{hist}^{(i)}(.,g)=\sum_{k}\sum_{j\in B(k,g)} x_{C,j}^{(i)} / \sum_{k}\sum_{j\in B(k,g)} x_{P,j}^{(i)},$$

where *B(k,g)* is the set of reporting units contained in the summary cell for NAICS code *k* and geographical identifier *g* in both the current and prior census. This average historic ratio across NAICS codes was then used to calculate a transformed historic ratio

$$R'^{(i)}_{hist}(k,g) = R^{(i)}_{hist}(k,g) \big/ \overline{R}^{(i)}_{hist}(.,g) .$$

One of the two score values calculated for each historic cell ratio was used to detect extreme values across NAICS codes for each geographical identifier $g\in G$:

$$z_1(k,g) = | R'_{hist}(k,g) - 1 | / SD(R'_{hist}(k,g); over\ k),$$

where *SD(u; over v)* denotes the standard deviation of *u* over *v*. The other score value calculated for each historic cell ratio was used to detect extreme values across geographical identifiers $g\in G$ for each NAICS code *k*:

$$z_2(k,g) = \frac{|R_{hist}(k,g) - \overline{R}_{hist}(k,.)|}{SD(R_{hist}(k,g); over\ g \in G)} .$$

Similarly a score value, $z_{3,}$ was calculated for each current cell ratio and used to detect extreme values of current cell ratios across geographical identifiers $g\in$ G for each NAICS code k. Current cell ratios varied too much among NAICS codes to make across NAICS code comparisons meaningful.

Batch processing was used to calculate the three *z*-score values for every cell. Cells with one or more z scores greater than 1.78 were labelled *initial outliers*. It was necessary for several other conditions to be satisfied in order for an initial outlier to be labelled a *final outlier*. One of these conditions was that *very small cells* could not be labelled as final outliers. These were cells in which all cell values were less than specified cut-offs, which varied across sectors. For example, in the Wholesale sector the very-small-cell cut-off for the number of reporting units was 10 and for the cell's total number of employees was 20. Final-outlier cells also had to have one or more *z* scores that exceeded a cut-off value that was much larger than 1.78. This higher cut-off value was set by supervisory staff based on their examination of the distribution of *z* scores and their estimate of how many outlier cells it would be possible to review with available resources.

After the batch processing identified the final outlier cells, information about which cells had been labeled as outliers was loaded into an interactive system for reviewing Economic Census data. Subject matter experts could then perform interactive queries to identify reporting units in outlier cells. The resulting lists of reporting units could be sorted by cell values to identify influential reporting units.

2.2 1997 Economic Census

Hogan (1995) describes how between the times of the 1992 and 1997 Economic Censuses analysts started using exploratory data analysis methods to detect outliers in data from individual reporting units for a number of Census Bureau's monthly and annual economic surveys. These analysts used SAS/INSIGHT[1] software to display scatter plots and box-and-whisker plots of their data. Subject matter experts identified outliers visually in these plots, and by clicking with their mouse on a plotted data point they were able to produce a tabular display of all the data fields for the plotted point's associated reporting unit.

For the 1997 Economic Census, three cell-outlier-review systems were developed that calculated historic and current cell outliers and displayed these ratios in

[1] SAS, SAS/INSIGHT, PC/SAS, BASE/SAS and SAS/CONNECT are registered trademarks of SAS Institute Inc.

SAS/INSIGHT graphs, which subject-matter experts used to visually identify cell outliers and to obtain additional information about the associated reporting units. The first system was used prior to the release of the Advance report to detect state-level outliers for NAICS codes representing sub-sectors and major industry groups. This system was not able to "drill down" from the cells to the reporting-unit data, so once a state was identified as an outlier separate listings of the largest data values in the state were reviewed for needed corrections.

A second cell-level-analysis system was developed to perform the analysis of state, county, and place outliers needed for pre-publication review of the GAS reports. This second system "required too much interactive time and effort by the analysts to locate outliers" (Lee, 2000), and it also lacked drill-down capability. This second system was not widely used. A third cell-level-analysis system was then developed, which was integrated with the interactive system that subject matter experts used to search for cell-level records and reporting-unit records satisfying entered queries. To use this cell-level-analysis system, subject matter experts first entered a query defining the set of NAICS codes and geographical identifiers to be analyzed. Cell-level records satisfying this query were displayed in a table and exported to an Excel spreadsheet, which PC/SAS read and provided to SAS/INSIGHT for the display of plots of historic and current cell ratios. When a point was clicked on in a plot, its observation number was displayed. This observation number could then be clicked on in the table of cell-level records to drill down to associated reporting unit records. This third system "required too much interactive time and effort by the analysts to locate outliers. [There were] too many graphs to review and too many steps involved to get to the SAS graph" (Lee, 2000). This third system was also not widely used.

3. Outlier Cell Analysis for 2002 Economic Census

At the beginning of our planning for outlier cell analysis for the 2002 Economic Census, we decided to return to the approach of the 1992 Economic Census in which batch processing, not interactive analysis, identified the cell outliers. We also decided there would an interactive system, which would permit subject-matter experts to select a set of outlier cells and then drill-down to associated high-influence reporting units.

3.1 Detection of Cell Outliers

Between the 1992 Economic Census and the 2002 Economic Census, the Census Bureau began using resistant methods to determine editing parameters for Economic Census data (Thompson and Sigman, 1999, and Thompson, 1999). This work plus the various macro-editing approaches described in Granquist (1991) prompted us to perform a small comparison study to determine how outlier cells should be identified for the 2002 Economic Census. One of the methods we evaluated was the Hidiroglou and Berthelot (1986) edit which calculates an outlier score by performing a series of transformations on a historic cell ratio or a current cell ratio. The first transformation, denoted *S*, is

$$SR = \begin{cases} R/R_m - 1 & R \geq R_m \\ 1 - R_m/R & 0 < R \leq R_m \end{cases}$$

where *R* is either a historic cell ratio or a current cell ratio, and R_m is the median of R over a set $\mathbb{R}$ (discussed in more detail later in this section). The second transformation, denoted *E*, is for historic ratios

$$ESR(u) = (SR)[\max(T_P, T_c)]^u,$$

where 0<u<1, and for current ratios

$$ESR_{curr}^{(i,i')}(u) = (SR_{curr}^{(i,i')})\left[\max\left(T_C^{(i)}, R_{curr,m}^{(i,i')}\ T_C^{(i')}\right)\right]^u,$$

where $R_{curr,m}^{(i,i')}$ is the median of $R_{curr}^{(i,i')}$ over $\mathbb{R}$.

According to Hidiroglou and Berthelot (1986, p. 76) the "exponent u in [this] transformation provides a control on the importance associated with the magnitude of the data. This transformation allows us to place more importance on a small change associated with a 'large' unit as opposed to a large change associated with a 'small' unit." The third transformation, denoted *Q*, is the quartile-transformation, defined generically as

$$QV = \begin{cases} (V - V_m) / D_{V,Q3} & V \geq V_m \\ (V_m - V) / D_{V,Q1} & V \leq V_m \end{cases}$$

V_{Q1}, V_m, V_{Q3} = first quartile, median and third quartile, respectively, of the *V*'s

$D_{V,Q3} = max\{V_{Q3} - V_m, |A*V_m|\}$

$D_{V,Q1} = max\{V_m - V_{Q1}, |A*V_m|\}$,

where *V* is the value of raw or transformed data. According to Hidiroglou and Berthelot, the purpose of the , $|A*V_m|$ term is to avoid problems when $V_{Q3} - V_m$ or $V_m - V_{Q1}$ is very small. They suggest the use of *A*=0.05. For the developmental testing and production processing for the 2002 Economic Census, we used *A*=0. Changing to a value of *A*>0.0 should be considered as a possible improvement to the outlier processing in future economic censuses.

As part of the planning for the 2002 Economic Census, we worked with several subject-matter experts to evaluate *QESR(0.3)*, *QESR(0.5)*, and *QESR(0.7)* along with the following four additional outlier scores:

- R_{hist}
- Z_2 = resistant version of z_2 = 10% trimmed t-test statistic (i.e. $\overline{R}_{hist}(k,.)$ is a 10% trimmed mean and *SD()* is a 10% Winsorized standard deviation)
- *QR*= quartile transformation of R_{hist}
- *QSR=QESR(0)* = quartile transformation of *SR*

For the R_{hist} score, we used *outlier thresholds* of 0.10 and 10; that is, a cell value was labelled an outlier if R_{hist}<0.10 or R_{hist}>10. For Z_2, we used outlier thresholds of +/- 3. For the quartile-transformation scores, we used outlier thresholds of +/- 4. These correspond to the inner-fence values in a Tukey (1977, p. 44) box-plot when $d_{Q1}=d_{Q3}=(Q3-Q1)/2$.

We compared the seven outlier scores on two data sets from which state-level cell totals for every state in the U.S. were calculated and three data sets from which county-level cell totals for every county in Texas were calculated. Each data set was for a different sector and contained data for the variables published in the sector's GAS reports for between four and six of the sector's (six-digit NAICS) industries. The source of x_p was final data from the 1992 Economic Census, and the source of x_c was preliminary data from the 1997 Economic Census.

We developed a BASE/SAS program that called PROC INSIGHT to interactively display seven scatter plots of R_{hist} versus T_c. For the two data sets used to calculate state-level cell totals, each plotted point represented a state, and for the three data sets used to calculate county level cell totals, each plotted point represented a county in Texas. The set of plotted points for all seven displayed plots were the same—state or county cell totals for a particular data item and a particular industry. The difference between the seven plots was that different outlier scores were used to indicate which cells were outliers. (The set $\mathbb{R}$ was all states in the U.S. for state-level outliers and all counties in Texas for county-level outliers.) Outliers were plotted in red and non-outliers were plotted in black. Subject-matter experts from each of the five sectors used the BASE/SAS program to select their preferred outlier scoring method

The BASE/SAS program allowed users to click their mouse on a plotted point to obtain additional cell-level information--such as the number of reporting units in a cell or the value of T_p for a cell. The program also allowed them to delete plots from the display so that non-preferred outlier scoring methods could be eliminated to make it easier to select the preferred outlier scoring method on the remaining plots. Each subject-matter expert (or team of subject-matter experts) repeated the process of displaying the seven plots and then selecting a preferred outlier scoring method for each data item and industry in a particular sector.

We had hoped the results of the comparison study would indicate one outlier scoring method was preferred across all five sectors. This was not the case. Two sectors had as their highest preference the *QESR(0.3)* method, another two sectors had as their highest preference the Z_2 method, and one sector had as their highest preference the *QR* method. The subject-matter experts in the two sectors that preferred the Z_2 method had prior experience with the 1992 outlier analysis system, which also used *z* scores to identify outliers, and this may have influenced their comparison of scoring methods. The other three sectors did not have prior experience with using the 1992 outlier analysis system. By examining the second-highest preferences across the five sectors, it was decided that for the production processing, two outlier scores would be used—*QESR(0.3)* and *QSR*—and that it would be necessary for a ratio to be labeled as an outlier by both scores in order for it to be included in the set of outliers to be reviewed.

When there is an odd number of ratios in $\mathbb{R}$, the median of *ESR* for $R\in\mathbb{R}$ is equal to zero. When there is an even number of ratios in $\mathbb{R}$, the median of *ESR* for $R\in\mathbb{R}$ is very close to zero, if not equal to zero. When the median of *ESR* for $R\in\mathbb{R}$ is equal to zero, the requirement that a ratio must be classified an outlier by both |*QESR(0.3)*| > *c* and |*QSR*| > *c* is equivalent to comparing the absolute value of the following composite score to the outlier cut-off *c:*

$$SCORE2=\begin{cases}\min(QESR(0.3),QSR) & R\geq R_m\\ \max(QESR(0.3),QSR) & R\leq R_m\end{cases}$$

When $max(T_p,T_c)$ is small, *SCORE2* equals *QESR(0.3)*; when $max(T_p,T_c)$ is large, *SCORE2* equals *QSR*. This satisfied study participants who felt *QESR(0.3)* identified too many outliers for large values of $max(T_p,T_c)$. Banim (2000) also compared different *u* values when using the Hidiroglou-Berthelot edit to macro-edit economic data and decided to use *u*=0.2.

In the production processing for the 2002 Economic Census, we calculated *QESR(0.3)* and *QSR* scores for both historic and current cell ratios for all the cell values. We classified a ratio as an outlier if (1) its cell values did not satisfy the very-small-cell conditions (for both 1997 and 2002 for historic ratios, for 2002 for current ratios) and (2) |*QESR(0.3)*|>4.0 and |*QSR*|>4.0. Not all the ratios classified as outliers were made available for additional analysis. For historic cell ratios, outliers were excluded from additional analysis if the number of reporting units contained in the cell was less than three in both 1997 and 2002. For current cell ratios, outliers were excluded from additional analysis if the number of reporting units contained the cell was less than three in 2002.

Some states have a small number of counties or have a small number of places with 2000 Census population greater than 2500. Also, some industries may have very few counties or places in a particular state with nonzero cell totals. Consequently, for the production processing for the 2002 Economic Census, we let $\mathbb{R}$ be the set of all counties in the U.S. for county-level cells and the set of all places in the U.S. for place-level cells. Selecting $\mathbb{R}$ appropriately is a possible area of research that could improve cell-outlier processing in future economic censuses.

3.2 Determination of High-influence Reporting Units

In early 2004 we modified the software developed for the comparison study to run in batch on mainframe Hewlett-Packard GS160 Alpha server running OpenVMS. In addition to identifying cell outliers, the batch software created approximately 4,000 small files, called *detail files*, containing data for up to 20 high-influence reporting units associated with each outlier cell value. For historic ratios we used the influence measure I_{hist} , defined by equation (1), to determine which reporting units would be the sources of the data contained in the detail files. For current ratios we used a modified form of I_{curr} obtained by replacing R_{curr} in equation (2) with $R_{curr,m}$, the median of R_{curr} over $\mathbb{R}$:

$$I_{curr}^{*(i,i')}(j^*) = (x_{C,j^*}^{(i)} - R_{curr,m}\, x_{C,j^*}^{(i')}) / T_C^{(i')}$$

This modification makes I^*_{curr} a measure of *cell contribution to cell extremeness*, which we discuss in more detail below. For *positive outliers* (i.e., $R>R_m$), reporting units with the largest (positive) values of I_{hist} or I^*_{curr} in a cell were the source of detail-file data; whereas for *negative outliers* ((i.e., $R < R_m$), reporting units with the smallest (negative) values of I_{hist} or I^*_{curr} in a cell were the sources of detail-file data.

We decided to organize the detail files as a large number of small files, so they could be quickly accessed and displayed by interactive client-server software. The client process was a SAS/AF application launched from PC/SAS that used SAS/CONNECT to obtain needed data from the mainframe computer.

The first screen of the SAS/AF application was a menu screen. The user selected a set of outliers to analyze by specifying the following information:

- One of the 15 service-economy sectors
- State, county, or place-level outliers
- Historic or current cell outliers
- One of two sets of cell values: core values (sales, number of employees, and first-quarter or annual payroll) or other values
- A state for county- and place-level outliers, or a set of states for state-level outliers
- A set of NAICS codes

The second screen of the SAS/AF application displayed high-influence reporting-unit-level data for the selected set of outliers. This screen also displayed *SCORE2* and information indicating the contribution of reporting-unit-level data to some measure of extremeness of the associated cell ratio. For historic ratios, the indication of cell-value extremeness was $100^*(R_{hist}\text{-}1)$, the percentage change of the 2002 cell total from the 1997 cell total. The reporting-unit contribution measure was 100^*I_{hist}, which was labelled as the reporting-unit's contribution to the cell's percentage change. For current ratios the values of R_{curr} and $R_{curr,m}$ were displayed next to each other, and the reporting-unit contribution measure was

$$100^*(I^*_{curr})/(R_{curr} - R_{curr,m}),$$

which was labelled as the reporting unit's contribution, expressed as a percentage, to the difference between the cell's current ratio and the median current ratio.

Users navigated to the second screen of the SAS/AF application by clicking on a "Get Report" button on the first screen. When the second screen came up, the reporting-unit-level data was sorted in a default order. The second screen provided a capability to temporarily change the order of sorting. The default sort was different for state-, county-, and place-level outliers. For state-level outliers, the first variable in the sort key was the cell NAICS code and the second-to-last variable in the sort key was the type of cell value (sales, annual payroll, etc). The advantage of this sort order was that all extreme cell values for a particular reporting unit were displayed near to each other, making it possible for subject-matter experts to simultaneously investigate possible reporting or processing errors for multiple data variables in a single reporting unit. For county- and place-level outliers, the first variable in the sort key was *SCORE2*. The advantage of this sort order was that it allowed subject-matter experts to further subset the set of outliers they wanted to analyze with an outlier cutoff equal to some value greater than 4.0 or less than -4.0. The disadvantage of this sort order was it could be difficult to see easily all extreme cell values associated with a particular reporting unit.

3.3. Usage Statistics

The GAS reports are published on a flow basis. For the 2002 Economic Census, a small number of GAS reports were published in late 2004, with the majority of the GAS reports published in the first half of 2005. As of April 29, 2005, the SAS/AF program for selecting outliers and displaying reporting-unit-level data had been used by 80 subject matter experts. Table 1 displays the monthly number of users and logins from October 1, 2004, though April 29, 2005.

Forty-one subject matter experts used the software during March 2005. Twenty-four users (58%) used the software between one and five times during the month; twelve users (29%), between six and 10 times; four users (17%) between 11 and 20 times; and one user (5%) used the software more than 20 times during the month.

3.4. User Feedback

In February 2005, a questionnaire about the cell outlier analysis system was sent to 70 subject-matter experts involved in the preparation of the GAS reports. Twenty questionnaires were returned, with 18 respondents replying they had used the system. The summary by Shoemaker (2005) of the data from these 18 respondents contains the following information:

- Sixteen (89%) of the respondents found the speed and performance of the system acceptable.
- The general response to the survey indicated users very satisfied with the system. In the four questions about how the system performed, 101 of 122 responses were positive. The main area of negative comments concerned the presentation of unnecessary information that resulted from including classification and data items that only pertained to selected sectors, which was the result of the development of a generalized system.
- Question 21 of the questionnaire asked respondents to “Estimate the percent[age] of outliers cells that you reviewed that fell into each of the following categories:”, which was followed by descriptions for three categories. Table 2 contains the category descriptions and the average of the responses.

3.5. Data-based Evaluation

On September 27, 2004, we archived a copy of the data files that contained state-, county-, and place-level cell totals, outlier flags, and outlier scores for four sectors: Wholesale, Retail, Real Estate, and Healthcare. Subject-matter experts who reviewed Real Estate sector data started using the cell outlier analysis system in late 2004, and in December 2004 they published GAS reports for four states: Alaska, Hawaii, Maine, and Montana. By comparing the archived cell totals, which we refer to as *preliminary* values, to cell totals calculated from data obtained after the release of the associated GAS report, which we refer to as *final* data values, we were able to analyze if preliminary cell totals labelled as outliers are indeed *bad*—i.e., their preliminary value differs from their final value—and, conversely, if preliminary cell totals that are bad are being identified as outliers. Some of our analysis findings were the following:

- Nearly all the cells identified as outliers were bad, but only a small proportion of bad cells were identified as outliers;
- More outliers were identified by current-cell ratio than by historic-cell ratios; and
- Simultaneous analysis of historic- and current-cell ratios identifies more outliers than separate analyses.

Additional details about this data-based analysis are available from the author.

4. Possible Improvements and Additional Research

Two candidate improvements have already been mentioned:

- Using a value of $A>0$, e.g. $A=0.05$, instead of A=0
- Making $\mathbb{R}$ all the counties or places in a state or a group of states for county- and place-level outliers, instead of all the counties or places in the U.S.

Some additional candidate improvements are the following:

- Providing a both-cell-ratio analysis option, instead of the current separate analyses of historic and current ratios
- Development of a data-based approach for selecting numerator and denominator variables for current cell ratios
- Instead of sorting the reporting-unit display for county- and place-level cells by SCORE2, which nearly always has different values for different cell values in the same cell, sort the display by say SCORE3, which is some summary measure (e.g., minimum, maximum, average, etc) of SCORE2 over the cell values in the same cell. This would make it easier for subject matter experts to simultaneously review all of the cell values in the same cell when a cell has multiple extreme cell ratios.
- When one reporting unit in an outlier cell has much greater influence than other reporting units it is very possible this reporting unit contains a reporting or processing error. Hence, it may be

beneficial to incorporate information about the distribution of very high influence reporting units within a cell into the cell-outlier detection process.

Acknowledgements

The author would like to thank Jock Black, Carma Hogue, Fane McLaughlin, Rita Petroni, Mike Shimberg, Dennis Shoemaker, and Katherine Thompson of the U.S. Census Bureau for their helpful comments on earlier versions of this paper.

References

Banim, J (2000). "An Assessment of Macro Editing Methods," UN/ECE Work Session on Statistical Data Editing, Cardiff, United Kingdom, 18-20 October, 2000. www.unece.org/stats/documents/2000/10/sde/7.e.pdf

Braam, T. (1992). "Outlier Specification: Standard Ratios and Z-Scores," unpublished memorandum, U.S. Census Bureau, 92EAG-B-G/17/E/1, Supplement A, September 10, 1992, revised December 16, 1993.

Granquist, L. (1991). "Macro-editing: Methods for Rationalizing the Editing of Quantitative Data ," monograph paper presented at the International Statistical Seminar, Basque Statistical Institute, Vol. 24, www.eustat.es/prodserv/datos/vol0024.pdf. Also published in *Statistical Journal*, Vol. 8, pp. 137-154.

Hidirglou, M. and Bethelot, J. (1986). "Statistical Editing and Imputation for Periodic Business Surveys," *Survey Methodology*, Vol. 12, pp. 73-83.

Hogan, H. (1995). "How Exploratory Data Analysis is Improving the Way We Collect Business Statistics," Proceedings of the Survey Research Section, American Statistical Association.

Lee, P. (2000). "1997 Outliers System," unpublished memorandum, U.S. Census Bureau, June 27, 2000.

Mesenbourg, T., Walker, E., and Hanczaryk, P. (2003). "The Census Bureau's Business Register: Basic Features and Future Direction," Joint UN/ECE/Eurostat Seminar on Business Registers, Luxembourg, 25-26 June 2003. www.uncec.org/stats/documents/ces/sem.50/5.e.pdf .

Shoemaker, D. (1993). "Outlier Specification," unpublished memorandum, U.S. Census Bureau, 92EAG-B-G/17/E/1, Revision 3, April 15, 1993.

Shoemaker, D. (2005). "Outlier User Report Summary," unpublished memorandum, U.S. Census Bureau. May 2005.

Thompson, K (1999). "Ratio Edit Tolerance Development Using Variations of Exploratory Data Analysis (EDA) Resistant Fences Methods. Statistical Policy Working Paper 29, available from the Federal Committee on Statistical Methodology www.fcsm.gov/99papers/thompson.pdf

Thompson, K. and Sigman, R. (1999). "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data," *Journal of Official Statistics,* Vol. 15, pp. 517-535.

Tukey, J. (1977). *Exploratory Data Analysis*, Addison-Wesley.

Table 1. Monthly number of users and logins.

	Oct '04	Nov '04	Dec '04	Jan '05	Feb '05	Mar '05	Apr '05
# unique users	17	38	44	49	49	41	34
# logins	93	137	277	432	278	271	181

Table 2. Question 21 category descriptions and average responses

Category Description	Average response (n=15)
21(a). Cells that require correction	27%
21(b). Significant (i.e. extreme) cells that required review but not correction	47%
21(c). Cells that the outlier system should not have identified	26%