DRAFT
8/23/2007

An Empirical Investigation into Macro Editing

Katherine J. Thompson[1] and Laura Ozcoskun
Office of Statistical Methods and Research for Economic Programs, U.S. Census Bureau

Introduction

Before releasing estimates for publication, most period surveys systematically review their computed estimates. Macro-editing is an identification process used to determine whether outlying estimates are the results of uncorrected respondent or data capture errors or are in fact values that provide useful information (e.g., indicators of change in target estimates). Such identification is generally performed after completing micro-level review, during which the individual questionnaire returns are scrutinized and corrected on a flow basis. At the macro-level review phase, **distributions** of tabulated cell estimates are reviewed, within both the current collection period and in contrast to corresponding prior period estimates. Once outlying cell estimates are flagged, analysts review the micro-data within these cells -- often focusing on the most influential reporting units -- and make any **necessary** corrections.

The majority of the statistical macro-editing techniques rely on distributional analyses, attempting to isolate atypical estimates. Survey data estimates rarely have known parametric distributions. Moreover, quantitative economic data are often best assessed via ratio comparisons of totals (e.g. current to prior estimates, wage per employee). Consequently, macro-editing techniques that utilize survey data must employ non-parametric or robust methods. Moreover, since the original set of estimates will contain outliers, these methods should be resistant. As always with data sets containing multiple outliers, macro-editing may be subject to two types of outlier-identification problems: 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.

Ratio comparisons are often quite effective at identifying outlying estimates, but can lead to redundant analyst work since often the same estimation cells are repeatedly identified using different sets of estimates. A multivariate outlier detection method that simultaneously considers all key estimates to identify all (or most) outlying estimation cells could save considerable time. Using data from the 2002 and 2003 data collections of the estimates collected from the U.S. Census Bureau's Annual Capital Expenditures Survey (ACES), Thompson (2007) presented promising preliminary results for bivariate comparisons with applications of the Hidiroglou-Berthelot edit (Hidiroglou and Berthelot, 1986) and with resistant fences methods, and had some success with multivariate outlier detection methods by applying a robust Mahalonobis distance measure.

This paper continues the evaluation presented in the earlier paper, by applying these recommended techniques to subsequently collected ACES data in addition to data from two different economic programs administered by the U.S. Census Bureau. The objective of this study was to determine whether any of the previously recommended methods could be successfully utilized with few modifications by other programs. If so, these methods could be incorporated into our directorate's Standard Economic Processing System (StEPS). Analysts could apply these methods to stored estimates as part of the review process. Note that we **do not** advocate any outlier-treatment procedure as a result of macro-editing. This analysis is designed simply as an additional stage of estimate validation, for the purposes of analyst identification (and possible review of micro-data).

In the next section, we present the methodology used, including the outlier detection methods, the evaluation procedure, and the evaluation methodology. We then present case studies from three periodic economic surveys: one survey that employs a very typical design and estimation procedure, but collects fairly atypical economic data (ACES); one survey that collects representative economic data, but utilizes a less typical design strategy and employs a very atypical estimation procedure (the Quarterly Financial

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Report, or QFR); and one survey that has a very conventional design and estimation procedure and collects representative economic data (the Quarterly Services Survey, or QSS). Finally, we make a few concluding comments and recommendations for future research.

Methodology

Outlier Detection Methods

Bivariate (Ratio) Comparison Methods. Bivariate macro-level analyses perform two types of (estimate) comparisons: ***current cell ratio tests*** and ***historic cell ratio tests***. Current cell ratio tests are designed to detect extreme observations within the current collection period in the context of the entire survey, by comparing two different item estimates from the **same data set** (current period data) in the same estimation cells. More formally, let $\hat{Y}_{C,i}$ and $\hat{X}_{C,i}$ be the survey estimates of two highly correlated items from estimation cell *i*, both collected at current time *t*. The current cell ratio $\hat{R}_{C,i}$ is the quotient of $\hat{Y}_{C,i}$ and $\hat{X}_{C,i}$, and the complete set of these current cell ratios is denoted $\hat{\mathbf{R}}_C = \{\hat{R}_{C,i}, \forall i\}$. Historic cell ratio tests are designed to detect extreme fluctuations in corresponding survey estimates between consecutive time periods, formally comparing the value of an estimate $\hat{Y}_i^t$ ($= \hat{Y}_{C,i}$ as defined above) to its corresponding prior period value $\hat{Y}_i^{t-1}$. The complete set of historic cell ratios for item *P* is denoted $\hat{\mathbf{R}}_P = \{\hat{R}_{P,i}, \forall i\}$.

A traditional "ratio edit" compares the ratio of two items to predetermined edit limits (tolerances). This type of edit can be quite problematic with periodically collected data. For example, when the distribution of current cell or historic cell ratios is very positively skewed, then outliers on the left tail of the distribution are often undetectable with traditional robust (or even resistant) outlier detection methods. Equally problematic, unless ratio edit tolerances are developed within some type of unit-size classification, the variability of the ratios can be quite large, and the tolerances will need to be accordingly wide. When this happens, too many small estimates are erroneously flagged as outliers, and not enough large units will be considered. Hidiroglou and Berthelot (1986) refer to this as the "size masking effect" because the variability of ratios from smaller estimation cells is often legitimately larger than the variability of ratios from larger estimation cells.

Their Hidiroglou-Berthelot (HB) edit is an edit procedure that is specifically designed to address these two outlier-detection problems for distributions of ratios that are **strictly positive**. The HB edit performs the following series of transformations on the **original** distribution of current or historic cell ratios prior to outlier identification:

- **Centering transformation** $s_i = \begin{cases} \dfrac{\hat{R}_i}{m} - 1 & \hat{R}_i \geq m \\ 1 - \dfrac{m}{\hat{R}_i} & 0 < \hat{R}_i < m \end{cases}$

  where *m* is the median of the ordered distribution of ratios as defined in Section 2.1.2. above.

- **Magnitude transformation** $E_i = \begin{cases} s_i \{MAX(\hat{Y}_{C,i}, \hat{\bar{R}}\hat{X}_{C,i})^U & \text{(current cell ratios)} \\ s_i \{MAX(\hat{Y}_i^t, \hat{Y}_i^{t-1})^U & \text{(historic cell ratios)} \end{cases}$

  where $\hat{\bar{R}}$ is an industry-average or median current cell ratio[2] and $0 \leq U \leq 1$. The industry-average ratio ensures that both of the two current cell ratio items are converted to the same units of measure (e.g., a wage per employee ratio).

[2] In our applications, $\hat{\bar{R}} = m$.

The exponent U "provides control on the importance associated with the magnitude of the data" (Hidiroglou and Berthelot, 1986). For example, $U \geq 0.5$ will greatly compress large values of ratios (generally obtained from smaller units) and will leave smaller values of ratios virtually unchanged. Following the recommendations of Thompson (2007), Sigman (2005), and Banim (2000), we consider values of U = 0.30 and 0.50. Outliers are identified as values smaller than ($E_m - Cd_{Q1}$) or larger than ($E_m + Cd_{Q3}$), where E_m is the median value of the E_i, C is a parameter that controls the width of the acceptance interval (obtained subjectively, through trial and error), $d_{Q1} = MAX(E_m - E_{Q1}, |AE_m|)$ and $d_{Q3} = MAX(E_{Q3} - E_m, |AE_m|)$. The A parameter in the d_{Q1} and d_{Q3} terms avoids nearly-zero limits when the absolute distance from the E_m to E_{Q1} (the first quartile of the transformed ratios) or from E_m to E_{Q3} (the third quartile of the transformed ratios) is quite small. We used A =0.05 in the applications described in Section 3 and consider C=10 and C=20 in all applications.

The HB edit requires that both estimates in the ratio edit are positive. This requirement is often satisfied with economic data items such as sales/receipts or expenditures. As an alternative outlier-detection method for real-valued estimates, we consider asymmetric fences rules (Thompson, 1999). Given an ordered distribution of current or historic cell ratios, let q_{25} = the first quartile, q_{75} = the third quartile, m = the median, and H = ($q_{75} - q_{25}$), the interquartile range. Asymmetric Fences rules flag ratios less than $q_{25} - k \times (m - q_{25})$ or greater than $q_{75} + k \times (q_{75} - m)$ as outliers, elongating the outlier-cutoff rule in the direction of the distribution's longer tail: k =3 defines inner fences and k = 6 defines outer fences (c.f., k to the C parameter in the HB edit). Since they are based on quartiles, resistant fences rules are designed to reduce masking; the statistician controls the swamping via the number of interquartile ranges between the quartiles and the fences. Asymmetric fences rules are applied to the original (untransformed) distributions of ratios. When distribution of ratios has wide variation (e.g., both items are not highly correlated), then the asymmetric fences rules will be quite prone to the "size-masking effect" mentioned above.

Multivariate Outlier Detection Methods. With multivariate outlier detection, we consider p variables jointly to identify estimation cells that have outliers in **several** different variables. Graphically, an ellipse is placed around the data and the estimation cells that fall outside of the ellipse are flagged as outliers. The Mahalanobis distance statistic is the classical method used to identify outliers in R^p from a randomly sampled dataset ($\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$), where $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$, assumed drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, estimated respectively by T(**X**) and C(**X**). The Mahalanobis distance for each observation is given as $MD_i = (x_i - T(\mathbf{X}))C(\mathbf{X})(x_i - T(\mathbf{X}))'$. The distance MD_i measures how far each x_i is from the center of the cloud of data, while taking into account the cloud's shape. Outlying observations are identified by comparing the computed MD_i to a χ^2_p critical value.

The assumption of multivariate normality with economic micro-data is somewhat questionable, although multivariate lognormality is fairly common when all estimates are strictly positive. For macro-editing, we compare distributions of sample **estimates**, each in itself a function of a sample mean and can very loosely invoke a Central Limit Theorem (then cross our fingers) to assume multivariate normality.

The classical Mahalanobis distance is prone to masking effects because of the weak resistance from the parametric estimators for T(**X**) and C(**X**). We use the minimum volume ellipsoid (MVE) measure (Rouseeuw and Zomeren, 1990) to develop robust estimates of T(**X**) and C(**X**) with approximately 50% breakdown points. T(**X**) is determined from the center of the MVE covering half of the observations and C(**X**) is determined by the same ellipsoid after applying a correction factor. This method should have low incidence of masking because of their high breakdown points. Given the questionable assumption of normality in our case study data sets, we consider two variance stabilizing power transformations as well as applications to the original sets of estimates: the log-transformation[3] and the cube-root transformation.

[3] In two of our three case study data sets where all estimates are strictly positive.

Evaluation Methodology

Preliminary Analysis and Classification. Each case study uses estimates constructed from consecutive data collections in the same programs. The **input** data are constructed from weighted **reported** data, with edited values substituted for missing reported data items. As the first step of the evaluation, we classified each separate estimate in the **input** data set into the following categories by comparing the percentage difference of the input data estimate to its corresponding final data estimate. If the data item is strictly non-negative (e.g., capital expenditures, sales), then we classified each estimate as an **outlier** or **not an outlier** based on an analysis of each collection period's distribution of percentage difference for each item. Specific classification rules are included with each separate analysis. If the item could take on any real value (e.g., net income before taxes), then we classified each estimate as

Outlier (O)	The percentage difference between the input and final data estimates in this cell was greater than the robust confidence interval defined by $\left(\overline{R}_T \pm 2\sigma_W\right)$, where $\overline{R}_T$ is the 10-percent trimmed mean of the distribution of percentage differences and σ_W is the10-percent trimmed mean standard error (based on the 10-winsored sum of standard deviations).
Not an outlier (N)	The item in this cell was not flagged as an "O."

An obvious limitation of this classification procedure is the subjective determination of outlier "cut-off" values. This decision was data-based, varied by program, and is not meant as a recommendation for other data sets. Recall that the end-use of this evaluation is find outlier-detection methods, and we are **not** advocating any outlier-procedure, simply review of the micro-data.

The **final data** are constructed from the weighted final edited/corrected data[4]. The **input data estimates** are our analysis variables, and the **final data estimates** are our evaluation variables (i.e., the "gold standard" estimates). Each estimate includes survey-specific adjustments for unit non-response and outlier correction factors as appropriate[5].

For ***current cell ratio tests,*** all considered estimates are constructed from **input** data in the same collection period. To obtain sufficiently large sets of outliers in our comparisons, we classified **input** estimates as:

- Bivariate: a ratio pair of estimates is an outlier if either the numerator or the denominator is flagged as an **outlier (O)** and is not considered an outlier otherwise.
- Multivariate: a set of industry estimates (within the same survey collection) is an outlier if at least one estimate is flagged as an **outlier (O)** and is not considered an outlier otherwise.

For ***historic cell ratio tests***, we compare the **input** data value in the most recent collection period to its corresponding **final** data value in the prior collection period. This mimics a production environment, where the current period's cell value would be edited in comparison with the presumably previously validated prior period's corresponding cell value. Here, we classified a historic cell ratio as an outlier if the numerator is either flagged as an **outlier (O).** Note that we do not consider multivariate applications to historic cell comparisons.

Evaluation Statistics. Any outlier-detection rule is a hypothesis test, where the null hypothesis is that none of the considered estimates is an outlier. Errors occur in either direction, so that we can define the **Type I error rate** for each outlier detection test as the proportion of **non-outlier** estimates that are flagged as outliers by a given procedure and the **Type II error rate** for each outlier-detection test as the proportion of **outlier** estimates that are not flagged as outliers by a given procedure.

---

[4] Input and final estimates include the same unit non-response adjustment procedures and micro-level outlier downweighting factors, as applicable

[5] ACES and QFR adjust estimates for unit non-response; QSS imputes complete records. ACES estimates include a micro-level outlier correction procedure.

With a bivariate comparison, the Type I and Type II error rates for individual tests are controlled by modifying test rule parameters (i.e., the C and U parameters in the HB edit and the k parameter with asymmetric fences rules), recognizing that a decrease in one error rate will result in an increase in the other. Granquist (1995) introduces the **hit rate** (the proportion of flagged estimates that are outliers) as a measure of the operational efficiency of a given outlier-detection rule. When data items are subjected to more than one bivariate edit, then the individual edit Type II error rates are a poor measure of the **unidentified** outliers in the completely reviewed data set (Thompson and Sigman, 1999). The all-item Type II error rate, defined as the proportion of **outlier** estimates that are **not** flagged as outliers by any ratio test, is a better measure. We compute all-item Type II error rates with respect to the complete set of tested items in a set of either current cell or historic cell ratios.

For multivariate analysis, the Type I error rate is the proportion of **non-outlier records** that are flagged as outliers (with respect to the total set of identified non-outliers). The Type II error rate is calculated similarly as with the bivariate tests, but the denominator is the set of **records** that contain at least one outlier estimate. In this setting, the hit rate is equivalent to (1 - Type II error rate), i.e., the power of the multivariate test. Similarly to the bivariate comparison, we compute the multivariate Type I and Type II error rates with respect to the considered estimates (which may not comprise the entire multivariate record).

Case Studies

Annual Capital Expenditures Survey (ACES)

Background. The ACES survey collects data about the nature and level of capital expenditures in non-farm businesses operating within the United States. Respondents report their expenditures for the calendar year in all subsidiaries and divisions for all operations within the United States. ACES respondents report total capital expenditures, as well expenditures on Structures and expenditures on Equipment, hereafter referred to as Total, Structures, and Equipment. All characteristics are further sub-classified by New/Used purchases (e.g., New Structures, Used Structures).

The ACES universe contains two sub-populations: employer companies and non-employer companies. Different forms are mailed to sample units depending on whether they are employer companies (ACE-1) or non-employer companies (ACE-2). New ACE-1 and ACE-2 samples are selected each year, both with stratified SRS-WOR designs. The ACE-1 sample comprises approximately seventy-five percent of the ACES sample (roughly 46,000 companies selected per year for ACE-1 and 15,000 for ACE-2). Responding firms account for approximately 88 percent of the total capital expenditures estimate. More details concerning the ACES survey design, methodology, and data limitations are available online at www.census.gov/csd/ace.

This paper examines data collected on the **ACE-1** form. For the ACE-1 component of the survey, each company is classified into **one** industry for stratification, and these industry strata are subdivided into certainty and non-certainty size strata, based on primary source of revenue (Stetser et al, 1998). Sampled units are asked to report their information by industry category for the industries in which the company participates. This type of survey is referred to in-house as a "roster" survey, where the number (roster) of industries for a given sample unit is unknown until reported. The roster data are tabulated by the sampled units' self-reported industries. The ACES collects company level and roster data. In our analysis, we consider only roster data items and only examine estimates of totals.

Because capital expenditures within the same company are generally characterized by low year-to-year correlation, historic cell ratio comparisons are ineffective for this survey at both the micro- and macro-levels. We concentrate instead on current cell ratio comparisons, examining estimates of **Total** Capital Expenditures, Capital Expenditures on **Structures**, Capital Expenditures on **New Structures**, Capital Expenditures on **Used Structures**, Capital Expenditures on **Equipment**, and Capital Expenditures on **New Equipment**. None of these estimates can take on negative values. A zero-valued cell estimate would be rare, but is possible. Our current cell ratio comparisons are Structures/Total (1); New Structures/Structures (2); New Structures/Used Structures (3); Equipment/Total (4); and New Equipment/Equipment (5). Thompson (2007) demonstrates poor correlation between ratio edit estimates in tests (1), (3), and (4), rendering the asymmetric fences methods inappropriate for these tests. Instead, we utilize the HB edit for

the bivariate comparisons, as recommended in Thompson (2007). For ACES, items whose percentage difference between the input and final data estimates in a cell were greater than the 95th percentile of the given distribution were classified as outliers. This appeared to be quite reasonable, as ACES micro-data undergo an outlier-detection and correction procedure prior to estimate evaluation, and the majority of the ACES micro-editing procedures make "small" changes to the record to satisfy additivity requirements. Appendix A presents the counts of estimate classifications for the ACES current cell ratio (bivariate) and multivariate comparisons respectively. Estimates are tabulated by the self-reported (roster) industry.

Current Cell Ratio Comparisons. Table 1 presents the evaluation statistics using the HB edit on the individual bivariate tests. When examining each **individual** test, it appears that using a conservative value of $C = 20$ best balances the individual Type I and Type II error rates in most cases, and that the "influence parameter" value (U) does not have a noticeable effect. The effectiveness of the actual tests seems to vary quite a bit by data collection, with the collective set of ratio edits being quite effective with the 2002 and 2003 data sets, but are considerably less so with the 2004 and 2005 data sets.

Table 1: Evaluation Statistics for Bivariate ACES Comparisons

		Structures/ Total			New Structures/ Structures			New Structures/ Used Structures			Equipment/ Total			New Equipment/ Equipment		
Year	HB Parameters	Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate
2002	U=0.8,=100.04	0.52	0.69	0.13	0.50	0.27	0.10	0.53	0.37	0.00	0.61	1.00	0.02	0.50	0.75	
	U=0.8,=200.00	0.57	1.00	0.11	0.50	0.32	0.09	0.53	0.39	0.00	0.61	1.00	0.00	0.75	1.00	
	U=0.6,=100.07	0.61	0.50	0.17	0.50	0.22	0.14	0.47	0.32	0.03	0.61	0.69	0.03	0.50	0.60	
	U=0.6,=200.00	0.65	1.00	0.12	0.50	0.29	0.10	0.53	0.37	0.00	0.70	1.00	0.01	0.58	0.83	
2003	U=0.8,=100.10	0.37	0.30	0.08	0.50	0.25	0.09	0.50	0.31	0.05	0.57	0.38	0.07	0.43	0.45	
	U=0.8,=200.03	0.30	0.64	0.06	0.25	0.56	0.04	0.44	0.64	0.02	0.46	0.78	0.03	0.45	0.60	
	U=0.6,=100.13	0.27	0.40	0.09	0.45	0.25		0.12	0.46	0.25	0.07	0.50	0.38	0.08	0.38	0.45
	U=0.6,=200.07	0.40	0.40	0.07	0.25	0.53	0.09	0.25	0.55	0.02	0.46	0.78	0.04	0.45	0.55	
2004	U=0.8,=100.05	0.43	0.57	0.11	0.42	0.33	0.10	0.10	0.43	0.02	0.50	0.70	0.01	0.50	0.86	
	U=0.8,=200.00	0.50	1.00	0.10	0.42	0.37	0.06	0.20	0.53	0.01	0.71	0.80	0.00	1.00	0.00	
	U=0.6,=100.06	0.50	0.50	0.12	0.42	0.32	0.13	0.00	0.38	0.03	0.50	0.64	0.01	0.42	0.88	
	U=0.6,=200.00	0.57	1.00	0.11	0.42	0.35	0.10	0.20	0.38	0.02	0.64	0.71	0.00	0.92	1.00	
2005	U=0.8,=100.07	0.54	0.33	0.17	0.20	0.32	0.10	0.46	0.37	0.08	0.43	0.38	0.06	0.40	0.47	
	U=0.8,=200.06	0.69	0.31	0.13	0.20	0.39	0.06	0.54	0.46	0.06	0.64	0.36	0.01	0.47	0.80	
	U=0.6,=10 0.1	0.38	0.33	0.17	0.13	0.34	0.10	0.38	0.40	0.12		0.43	0.30	0.08	0.33	0.43
	U=0.6,=200.03	0.62	0.38	0.13	0.2	0.39	0.05	0.38	0.57	0.07	0.50	0.37	0.04	0.40	0.60	

The influence parameter's effect is visible when examining all tests jointly with regards to the six tested items. Table 2 presents the All-Item Type II error rates (proportion of unidentified outliers in the data set). Here, the smaller value of the influence parameter ($U = 0.3$) combined with the less conservative critical value decision rule ($C = 0.10$) yields the lower all-item Type II error rates in all but the 2005 data set.

Table 2: All-Item Type II Error Rates (ACES)

Year	HB Parameters	Type II Error Rate	Year	HB Parameters	Type II Error Rate
2002	U=0.3, c =10	0.43	2004	U=0.3, c =10	0.37
	U=0.3, c =20	0.50		U=0.3, c =20	0.48
	U=0.5, c =10	0.45		U=0.5, c =10	0.37
	U=0.5, c =20	0.50		U=0.5, c =20	0.52
2003	U=0.3, c =10	0.30	2005	U=0.3, c =10	0.27
	U=0.3, c =20	0.33		U=0.3, c =20	0.30
	U=0.5, c =10	0.30		U=0.5, c =10	0.20
	U=0.5, c =20	0.33		U=0.5, c =20	0.27

Here, obtaining a reasonable level of All-Item Type II error rates using the HB edit requires five separate ratio tests. Table 3 presents Type I and Type II error rates for the multivariate outlier-detection procedure determined using the MVE technique, with records comprised of four estimates (New and Used Structures, New and Used Equipment). We consider three separate variations here, computing the robust Mahalanobis distance on the original data, log-transformed data, and cube-root transformed data.

Table 3: Multivariate Comparisons (ACES):
New and Used Structures, New and Used Equipment

Year	Transformation	Type I Error Rate	Type II Error Rate
2002	None	0.42	0.46
	Log	0.06	0.48
	Cube-Root	0.27	0.51
2003	None	0.4	0.15
	Log	0.12	0.25
	Cube-Root	0.22	0.16
2004	None	0.44	0.23
	Log	0.06	0.42
	Cube-Root	0.31	0.24
2005	None	0.39	0.04
	Log	0.06	0.69
	Cube-Root	0.23	0.27

Here, the Type I error rates obtained using the original or cube-root transformed data are unacceptably high, and the Type II error rates are equally poor. The results with the log-transformed estimates are more promising, with the caveat that their results are derived from smaller sets of ratios than with the other two transformations, since zero-value estimation cells must be excluded.

Quarterly Financial Report (QFR)

Background. The Quarterly Financial Report (QFR) is a sample survey of companies from the mining, wholesale trade, and retail trade sectors having total assets of $50 million or more, and from the manufacturing sector having total assets of $250 thousand or more. The QFR sample is divided into panels that are rotated into and out of the survey, and each non-certainty sampled company is interviewed for eight consecutive business quarters. For any given quarter, eight panels selected from up to three different frame years are in the survey. Each year, a new sample of corporate tax returns is selected from the most recent tax year data. This new sample is split into four panels. Each quarter, one of the four new panels is introduced, and the panel that has completed all eight interviews is dropped from the survey. This type of rotating panel design yields precise quarterly change estimates.

The sampling frame for the QFR survey comes from the file of United States Internal Revenue System (IRS) corporate tax returns. Every year, the Census Bureau receives a list of corporate tax returns for the

previous year from the IRS and classifies all the companies by reported industry (sample industry) and total assets. Companies that have total assets of $250 million or more are included with certainty and are in the survey indefinitely. The remaining companies are stratified within sample industry and are further stratified within sample industry code by size; the within-industry size strata are referred to as the asset classes. The mining, wholesale trade, and retail trade sectors have one non-certainty stratum per sample industry, whereas the manufacturing sector has five non-certainty strata per industry.

It is possible for a QFR company to conduct business in a different industry than indicated by the sampling frame. QFR estimates are tabulated by the company-reported industry, not the sample (frame) industry. The revised industry classification is referred to as the enumeration industry. Classification changes are determined when the report is returned. Estimates of quarterly totals are unweighted means multiplied by an estimate of population size for the enumeration industry/size-classification. This population estimate incorporates both industry changes (from the sampling frame) and the rotation scheme. The QFR variable-weight estimates are further adjusted for non-response in the enumerated industry and asset class, using unweighted inverse response rates as advocated by Vartivarian and Little (2002). More details concerning the QFR survey design, methodology, and data limitations are available online in the source and accuracy statement of any publication table (http://www.census.gov/csd/qfr/).

The QFR collects income statement and balance sheet data from each surveyed company. From this data, the QFR publishes several key economic statistics, including quarter-to-quarter percent change in sales as well as estimates of total. The key data items examined here are total sales, net income after taxes (NIAT), net income before taxes (NIBT), total assets, and (stockholders') equity. A zero-valued cell estimate would be rare, but is possible.

NIBT and NIAT are real-valued estimates. Consequently, we apply asymmetric resistant fences rules (instead of the HB edit) to current cell ratio comparisons of NIAT/sales, NIAT/assets, NIBT/assets, NIAT/equity, and NIBT/equity. Although these ratios are extremely important indicators to QFR data users, the estimates in each bivariate test are not strongly correlated. We apply the HB edit to historic cell ratios for sales, income, and assets considering data from all four quarters of the 2006 data collections and the last quarter of the 2005 data collection. For QFR, items whose percentage difference between the input and final data estimates in a cell were greater than the 95th percentile of the given distribution were classified as outliers. Appendix B presents the counts of estimate classifications for the QFR current cell ratio (bivariate) and multivariate comparisons respectively. Estimates are tabulated by enumerated industry.

The examined QFR estimates have a very low incidence of either bivariate or multivariate outliers, especially in the first three quarters (2005Q4, 2006Q1, and 2006Q2). With these quarters' estimates, the Type I error rates will be exaggerated, the hit rates will be nearly zero (since there are generally few or no outliers), and the Type II error rates will be zero as well.

Current Cell Ratio Comparisons. Table 4 presents the evaluation statistics using the asymmetric fences rules on the individual bivariate tests. The extremely low Type II error rates and hit rates in the first three quarters are an artifact of the unusually small number of outliers in the data sets.

Clearly, the asymmetric fences rules are less than optimal for outlier-detection with these sets of ratios. Recall that these ratio pair estimates are **not** highly correlated, and that the resultant tolerances can be quite wide. Moreover, the very small number of outlying observations in the first three quarters of data result in exaggerated error and hit rate effects. Given this, the outer fences rules appear to be not entirely unpromising for all ratio tests but the NIAT/sales test. Due to the small number of outliers in the QFR data, the all-item Type II error rates do not provide more information, ranging from 0 in 2006Q2 (no outliers) and from 0.60 to 1 in the other quarters. Except for 2006Q3, All-Item Type II error rates are equivalent for each asymmetric fences rule (inner, outer) within quarter: there is a slight improvement in the rates using inner fences rules with the 2006Q3 data (0.67 with inner fences compared to 0.78 with outer fences).

Table 4: Asymmetric Fences Results (QFR)

Collection Period	Fences	NIAT/ Sales			NIBT/ Equity			NIAT/ Equity			NIBT/ Assets			NIAT/ Assets		
		Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate	Type I Error Rate	Type II Error Rate	Hit Rate
2005Q4	Inner	0.34	1.00	0.00	0.08	0.00	0.00	0.14	0.00	0.00	0.08	0.00	0.00	0.03	0.00	0.00
	Outer	0.17	1.00	0.00	0.06	0.00	0.00	0.06	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
2006Q1	Inner	0.46	1.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.09	1.00	0.00	0.00	1.00	0.00
	Outer	0.31	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	1.00	0.00	0.00	1.00	0.00
2006Q2	Inner	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
	Outer	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2006Q3	Inner	0.23	0.40	0.30	0.11	0.88	0.25	0.14	0.86	0.20	0.00	0.88	1.00	0.00	1.00	0.00
	Outer	0.13	0.80	0.20	0.07	0.88	0.33	0.07	0.86	0.33	0.00	1.00	0.00	0.00	1.00	0.00
2006Q4	Inner	0.07	0.50	0.67	0.13	0.83	0.20	0.18	0.56	0.44	0.16	0.80	0.17	0.00	1.00	0.00
	Outer	0.03	0.63	0.75	0.10	1.00	0.00	0.04	0.78	0.67	0.06	1.00	0.00	0.00	1.00	0.00

Table 5 presents Type I and Type II error rates for the multivariate outlier-detection procedure determined using the MVE technique, with records comprised of the five estimates tested in the above bivariate comparisons. Since NIAT and NIBT are real-valued, we consider two separate variations (original data cube-root transformed data).

Table 5: Multivariate Comparisons (QFR)

Year	Transformation	Type I Error Rate	Type II Error Rate
2005Q4	None	0.29	0.00
	Cube-Root	0.07	1.00
2006Q1	None	0.24	0.50
	Cube-Root	0.26	1.00
2006Q2	None	0.22	0.00
	Cube-Root	0.27	0.00
2006Q3	None	0.3	0.67
	Cube-Root	0.12	0.86
2006Q4	None	0.41	0.45
	Cube-Root	0.18	1.00

With the QFR data, quite reasonable results are achieved by applying a robust Mahalanobis distance to the **original** data. In fact, the multivariate results represent a substantial improvements over the combined univariate results in terms of (All-Item) Type II error rates.

Historic Cell Ratio Comparisons. As mentioned above, QFR publishes estimates of change for key quarterly statistics. Analysts review the quarterly estimates prior to release. For our historic cell ratio comparisons, we apply the HB edit to estimates of sales, assets, and equity. Appendix B presents the counts of estimate classifications for the QFR historic cell ratio (bivariate).

Table 6 presents the Type I error rates for each of the bivariate tests. The Type II error rates and hit rates are not displayed: for change periods when the Type I error rate is zero, the Type II error rate is also zero and the hit rate is undefined (no outliers); for change periods when the Type I error rate is greater than zero (2006Q3 to 2006Q2 and 2006Q4 to 2006Q3), the Type II error rates are 100-percent and the hit rates are

zero. Due to the small number of outliers in these data sets, all sets of error rates are quite exaggerated, although there is some evidence that the HB edit is effectively identifying outliers (when they exist) with *U=0.5* and *C* = 20.

Table 6: Type I Error Rates for QFR Historic Ratios

Change Period	HB Parameters	Item		
		Sales	Assets	Equity
2006Q1 to 2005Q4	U=0.3, c =10	0.09	0.09	0.03
	U=0.3, c =20	0.06	0.03	0.03
	U=0.5, c =10	0.06	0.03	0.03
	U=0.5, c =20	0.03	0.03	0.03
2006Q2 to 2006Q1	U=0.3, c =10	0.06	0.09	0.06
	U=0.3, c =20	0.00	0.03	0.03
	U=0.5, c =10	0.00	0.03	0.04
	U=0.5, c =20	0.00	0.00	0.00
2006Q3 to 2006Q2	U=0.3, c =10	0.00	0.00	0.00
	U=0.3, c =20	0.00	0.00	0.00
	U=0.5, c =10	0.00	0.00	0.00
	U=0.5, c =20	0.00	0.00	0.00
2006Q4 to 2006Q3	U=0.3, c =10	0.00	0.00	0.00
	U=0.3, c =20	0.00	0.00	0.00
	U=0.5, c =10	0.00	0.00	0.00
	U=0.5, c =20	0.00	0.00	0.00

Industry estimates of assets and stockholders equity tend to be fairly constant between quarters, so the results displayed in Table 6 are not particularly surprising to our subject-matter experts – in fact, they are expected, and provide a bit more evidence of the effectiveness of this edit technique for change estimates of sales (which tend to be more variable).

Quarterly Services Survey (QSS)

Background. The Quarterly Services Survey is a voluntary economic indicator survey whose primary purpose is to provide timely estimates of quarterly receipts (published about 75 days after the end of the reference quarter) and early estimates of calendar year receipts for selected service sectors. Currently, the QSS covers the following North American Industry Classification System (NAICS) sectors: Information; Professional, Scientific, and Technical Services; Administrative and Support and Waste Management and Remediation Services; and Hospitals and Nursing and Residential Care Facilities.

Approximately 5,000 sampling units were selected for the initial QSS sample. Sample maintenance activities are performed each quarter. During this process, out-of-business units are identified and removed from mailing; and newly formed sampling units are identified, subjected to a two-phase sampling process, and selected units are added to the sample. The questionnaire for every NAICS code collects quarterly receipts, receipts by class of customer, and reporting period if the reported receipts are not for the calendar quarter. If a unit does not respond or does not report receipts, a value is imputed based on other survey data and administrative records. Further details about QSS are at http://www.census.gov/indicator/qss/qsstechdoc.pdf. QSS collects receipts from all six-digit industries, and collects receipts and expenses from the Hospitals and Nursing and Residential Care Facilities sector. Since the latter sector comprises only four six-digit industries, we confine our analysis to historic cell comparisons of receipts, which are always positively value. We apply the HB edit to historic cell ratios for receipts considering data from the 2005 QSS data collection. For QSS, items whose percentage difference between the input and final data estimates of receipts a cell were greater than the 90th percentile are classified as outliers. Appendix C presents the outlier counts for the QSS data.

Since we are confined to one analysis variable with QSS (other analysis variables are derived from edited totals or are only collected in one sector), we restrict our QSS analysis to a historical cell ratio comparison.

Historic Cell Ratio Comparison. Table 7 presents the Type I and Type II error rates for each historic cell ratio test along with its associated hit rate.

Table 7: Evaluation Statistics for QSS Historic Cell Ratios (90th Percentile Outlier Cut-off)

Change Period	HB Parameters	Type I Error Rate	Type II Error Rate	Hit Rate
2005Q4 to 2005Q3	U=0.3, c =10	0.00	1.00	0.00
	U=0.3, c =20	0.00	1.00	0.00
	U=0.5, c =10	0.00	1.00	0.00
	U=0.5, c =20	0.00	1.00	0.00
2005Q3 to 2005Q2	U=0.3, c =10	0.00	0.25	1.00
	U=0.3, c =20	0.00	0.75	1.00
	U=0.5, c =10	0.00	0.25	1.00
	U=0.5, c =20	0.00	0.50	1.00
2005Q2 to 2005Q1	U=0.3, c =10	0.00	0.00	1.00
	U=0.3, c =20	0.00	0.50	1.00
	U=0.5, c =10	0.00	0.00	1.00
	U=0.5, c =20	0.00	0.50	1.00

With these data, the evaluations statistics are so "ideal", in particular with values of c = 10 (the less conservative outlier detection rule), that we suspected that they were an artifact of our outlier classification procedure. Table 8 presents the same statistics obtained after reclassifying the individual estimates based on the 75th percentile as outlier cut-off (another data-based decision).

Table 8: Evaluation Statistics for QSS Historic Cell Ratios (75th Percentile Outlier Cut-off)

Change Period	HB Parameters	Type I Error Rate	Type II Error Rate	Hit Rate
2005Q4 to 2005Q3	U=0.3, c =10	0.00	1.00	0.00
	U=0.3, c =20	0.00	1.00	0.00
	U=0.5, c =10	0.00	1.00	0.00
	U=0.5, c =20	0.00	1.00	0.00
2005Q3 to 2005Q2	U=0.3, c =10	0.00	0.63	1.00
	U=0.3, c =20	0.00	0.88	1.00
	U=0.5, c =10	0.00	0.63	1.00
	U=0.5, c =20	0.00	0.75	1.00
2005Q2 to 2005Q1	U=0.3, c =10	0.00	0.50	1.00
	U=0.3, c =20	0.00	0.75	1.00
	U=0.5, c =10	0.00	0.50	1.00
	U=0.5, c =20	0.00	0.75	1.00

Again, no false outliers are detected. The effect on Type II error rates is apparent, and shows the same pattern. If there are in fact several outliers in the set of QSS estimates, then the HB edit with u = 0.3 or 0.5 And c=10 will clearly be very effective in correctly identifying outliers, but will not be completely sufficient for data review. In a production setting, however, it is more likely that the first comparison shown in Table 13 is realistic, since estimates will not be examined until micro-editing is complete, and large errors in the micro-data are corrected.

Discussion

In this paper, we applied a variety of previously-proven outlier detection techniques to sets of estimates from three very different economic surveys. With the ACES data, the degree of statistical association (correlation) between estimates is inconsistent from one collection period to another. With the QFR data, the statistical associations between items are more constant both within and between statistical periods, but they are quite low for all current cell ratios. With the QSS data, only one data item can be examined. Neither the QFR nor ACES estimates appear to be multivariate normally distributed, although it appears that the log-transformed ACES estimates satisfy this condition and that the QFR estimates are sufficiently "close to multivariate normal" to use the robust MVE Mahalanobis distance to identify multivariate outliers.

When the data items are positively valued and there is some statistical association between tested item pairs, the HB edit was generally very effective. Not surprisingly, the asymmetric fences methods did not fare so well when applied to the poorly-correlated QFR current cell ratio tests. The standardization procedure employed by the HB edit provides a clear advantage over the asymmetric fences applications. The current cell ratio results for QFR could possibly be improved by applying a power-transformation (e.g., the cube-root transformation) to the original ratios to reduce the effect of legitimate large tail values, then applying the asymmetric fences rules to the data.

In all of these case studies, the outlier-detection method does not use predetermined limits, but instead dynamically identifies outliers in the data set at hand. This is a tremendous advantage over many traditional micro-editing techniques. Here, however, each outlier-detection method requires "rules" for setting the limits, and these rules may very well differ for each comparison. Ultimately, flexibility will be key, since "rules" may need to be modified on a flow basis as a procedure identifies too many or two few outliers.

Despite the success of the HB edit with the ACES data current cell ratio tests, we believe that it is worth pursuing the use of the minimum volume ellipsoid (MVE) Mahalanobis distance measure to identify outliers. With the ACES current cell ratios, several different HB edit tests were required to achieve reasonable all-item Type II error rates. With both the ACES and QFR data, similar results were achieved in one-pass with this robust MVE Mahalanobis distance. Along the same lines, an automatic procedure that flags bivariate pairs via the HB edit tests, then unduplicates records could be equally more effective.

The results presented in this paper demonstrate the need for creativity and flexibility for successful macro-editing. Each presented technique had varying success within survey (between data collection periods) and between surveys. Even so, each case study presents methods that worked well for the studied program. As long as the subject-matter experts can spend sufficient preparatory time learning these methods and developing item-specific outlier rules, there is quite a bit to recommend further evaluations and perhaps even production implementation of these methods on our directorate's StEPS system.

Acknowledgements

The authors thank Rita Petroni and Mark Sands for their useful 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

Granquist, Leopold (1995). Improving the Traditional Editing Process. Business Survey Methods. New York: John Wiley and Sons, pp. 385-481.

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

Rouseeuw, P.J. and Van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, **85**, 00. 633-639.

Sigman, R.S. (2005). Statistical Methods Used to Detect Cell-Level and Respondent-Level Outliers in the 2002 Economic Census of the Services Sector. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Stetser, M.C., Goodloe, J.M., Sands, M.S. (1998). The Evolution of Survey Methodology For A Company Survey Of Capital Expenditures. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 511-516.

Thompson, K.J. (2007, to appear). Investigation of Macro Editing Techniques for Outlier Detection in Survey Data. *Proceedings of the Third International Conference on Establishment Surveys*, American Statistical Association.

Thompson, K.J. (1999). Ratio Edit Tolerance Development Using Variations Of Exploratory Data Analysis (EDA) Resistant Fences Methods. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, available at www.fcsm.gov.

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

Vartivarian, S. and Little, R.J. (2002). On the Formation of Weighting Adjustment Cells for Unit Non-response. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, pp. 3553-3558.

Bivariate Classifications for ACES

Year	Classification	Structures/ Total	New Structures/ Structures	New Structures/ Used Structures	Equipment/ Total	New Equipment/ Equipment
2002	Not Outlier	123	122	119	123	121
	Outlier	23	12	15	23	12
2003	Not Outlier	123	119	116	120	121
	Outlier	10	12	16	13	11
2004	Not Outlier	123	123	124	123	123
	Outlier	14	12	10	14	12
2005	Not Outlier	163	150	126	161	159
	Outlier	13	15	13	14	15

Multivariate Classifications for ACES

Year	Cell Classification	Count
2002	Not Outlier	127
	Outlier	21
2003	Not Outlier	129
	Outlier	8
2004	Not Outlier	138
	Outlier	7
2005	Not Outlier	171
	Outlier	4

Bivariate Comparisons for QFR Current Cell Ratios

	Classification	NIAT/ Sales	NIBT/ Equity	NIAT/ Equity	NIBT/ Assets	NIBT/ Assets
2005Q4	Not Outlier	35	36	36	36	36
	Outlier	1	0	0	0	0
2006Q1	Not Outlier	35	36	36	35	35
	Outlier	1	0	0	1	1
2006Q2	Not Outlier	36	36	36	36	36
	Outlier	0	0	0	0	0
2006Q3	Not Outlier	31	28	29	28	29
	Outlier	5	8	7	8	7
2006Q4	Not Outlier	29	31	28	32	29
	Outlier	8	6	9	5	8

Multivariate Comparisons (QFR): Sales, NIAT, NIBT, Equity, Assets

Collection Period	Classification	Counts	Collection Period	Classification	Counts
2005Q4	Not Outlier	35	2006Q3	Not Outlier	27
	Outlier	1		Outlier	9
2006Q1	Not Outlier	34	2006Q4	Not Outlier	27
	Outlier	2		Outlier	11
2006Q2	Not Outlier	36			
	Outlier	0			

Historic Cell Outlier Counts for QFR

Change Period	Classification	Item		
		Sales	Assets	Equity
2006Q1 to 2005Q4	Not Outlier	35	35	36
	Outlier	1	1	0
2006Q2 to 2006Q1	Not Outlier	36	36	36
	Outlier	0	0	0
2006Q3 to 2006Q2	Not Outlier	35	34	34
	Outlier	1	2	2
2006Q4 to 2006Q3	Not Outlier	35	36	35
	Outlier	2	1	2

Outlier Counts for QSS Data

Change Period	Classification	Item
		Quarterly Revenue
2005Q4 to 2005Q3	Not Outlier	29
	Outlier	3
2005Q3 to 2005Q2	Not Outlier	28
	Outlier	4
2005Q2 to 2005Q1	Not Outlier	28
	Outlier	4