Investigation of Macro Editing Techniques for Outlier Detection in Survey Data

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

1. Introduction

An important step in the evaluation of survey data estimates is the identification of macro-level outliers. This identification process is used to determine whether outlying preliminary estimates are the results of uncorrected respondent or data capture errors or are values that provide useful information (e.g., indicators of change in target estimates). Macro-editing is generally performed after the micro-level review phase, when the individual questionnaire returns are scrutinized and corrected on a flow basis. During the macro-level review phase, distributions of tabulated cell estimates are scrutinized, within both the current collection period and in contrast to corresponding prior period estimates.

Macro editing techniques rely on distributional analyses, attempting to isolate atypical data points (estimates) from the bulk of the observations. Survey data estimates rarely have known parametric distributions. Moreover, quantitative economic data is often best assessed via **ratio** comparisons of totals, further complicating any parametric analysis. Consequently, macro editing techniques that utilize survey data must employ non-parametric or robust methods.

Prior to macro-review, it is quite likely that a given set of survey estimates will contain multiple outliers. This can lead 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. To reduce the probability of these phenomena, I utilize outlier-resistant methods (Hoaglin et al, 1983); the resistance "breaks down" (exceeds the breakdown point) when the actual number of outliers exceeds the expected number of outliers in a given distribution

Ratio comparisons are generally quite successful at identifying outliers. In many cases, however, the same estimation cells are repeatedly identified using different sets of estimates (ratios). A multivariate outlier detection method that simultaneously considers key estimates that identifies all (or most) outlying estimation cells could save considerable analyst research and processing time. In this paper I investigate a variety of robust and resistant bivariate and multivariate methods for detecting macro-level outliers using survey estimates from the U.S. Census Bureau's Annual Capital Expenditures Survey (ACES).

2. Outlier Detection Methods

2.1. Bivariate Methods (Ratio Comparisons)

In general, the macro-level bivariate analyses perform two types of (estimate) comparisons: ***current cell ratios*** and ***historic cell ratios***. ***Current cell ratios*** 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. ***Historic cell ratios*** detect extreme fluctuations in corresponding survey estimates between consecutive time periods. Because of data limitations specific to ACES, only current cell applications are discussed in sections 2.1.1 through 2.1.3 below.

Both current cell ratios and historic cell ratios are ratio edits. 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. The usefulness of a ratio edit in detecting outliers is highly dependent on the strength of the statistical association between both items (Thompson, 1999). A ratio edit implies a no-intercept regression model, where the numerator is the dependent variable. If the regression model is given by $\hat{Y}_i = \beta\hat{X}_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, then the correlation between the two ratio edit items is the appropriate measure of the statistical association of the two items. More often

[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 author and not necessarily those of the U.S. Census Bureau.

with economic data, however, the error increases proportionally to the independent variable, so that a more appropriate regression model is often given by $\hat{Y}_i = \beta\hat{X}_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \hat{X}_i\sigma^2)$, and the model-$R^2$ is the appropriate measure of statistical association. Thompson et al (2001) found that the sample correlation will be higher than the corresponding model-R^2 value in the presence of heteroscedasticity, but can be used as a very crude (always overestimating) proxy.

Hidiroglou and Berthelot (1986) describe the masking effects present in the low correlation ratio edits. When the distribution of ratios is very positively skewed, then outliers on the left tail of the distribution are undetectable. Second, unless ratio edit tolerances are developed within some type of unit-size classification, then the variability of the ratios will be quite large, and the tolerances will need to be accordingly wide. Consequently, too many small units will be flagged as outliers, and not enough large units will be considered. The authors refer to this as the "size masking effect" because the variability of ratios from small sample units is often legitimately larger than the variability of ratios from large sample units. Their statistical edit is specifically designed to ameliorate this event.

2.1.1. Robust Regression

Bienias et al (1994) proposed using resistant regression methods for **micro-level** ratio comparisons. I take a similar approach using the macro-level estimates as independent and dependent variables but instead employ a different regression estimator, the Least Trimmed Squares robust regression estimation method (Rousseuw, 1984). This resistant regression procedure minimizes the **median** of the squared residuals instead of the sum of the squared residuals and has a 50-percent breakdown point. This method is fairly free of masking effects, but may be sensitive to swamping. For each bivariate comparison, I fit the unweighted regression model ($\hat{Y}_{C,i} = \beta\hat{X}_{C,i} + \varepsilon_i$) then performed a residual analysis. Any value where absolute residual is greater than $3\hat{\sigma}$ (the robust M.S.E) is flagged as an outlier.

2.1.2. Resistant Fences

Several programs in the Economic directorate of the U.S. Census Bureau use modifications of the Exploratory Data Analysis Resistant Fences methods to develop tolerances for micro-level ratio edits (e.g., Cornett et al (2006), Fescina et al (2004), Thompson et al (2001)). 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. **Resistant Fences** flags ratios less than $q_{25} - k\times H$ or greater than $q_{75} + k\times H$ as outliers. **Asymmetric Fences** flags ratios less than $q_{25} - k\times(m - q_{25})$ or greater than $q_{75} + k\times(q_{75} - m)$ as outliers.

The value of k determines the fence's "rule." For resistant fences, $k = 1.5$ defines inner fences, $k = 2$ defines middle fences, and $k = 3$ defines outer fences. The standard Tukey boxplot is constructed using resistant inner fences rules. For asymmetric fences, $k = 3$ defines inner fences, $k = 4$ defines middle fences, and $k = 6$ defines outer fences (Thompson, 1999). 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.

Resistant fences rules implicitly assume symmetry. When distributions of ratios are highly skewed, it can be helpful to symmetrize the original distributions of ratios with a power transformation such as the natural logarithm or the cube root prior to applying the resistant fences rules, then apply the inverse transformation to these fences to obtain the final tolerances (Thompson, 1999). Note that resistant fences methods do not work for bimodal distributions or with distributions that have a non-zero interquartile range.

In practice, resistant fences methods are generally used to develop a "fixed" set of tolerances for micro-edits, where tolerances must be developed from prior-period data so that questionnaires can be edited as they are received. In contrast, macro editing is dynamic, and new tolerances are developed for and from each studied data set.

2.1.3. Hidiroglou-Berthelot (HB) Edit

The Hidiroglou-Berthelot (HB) edit is a procedure originally designed to detect outlying historic cell ratio values in periodically collected micro-data. This method is employed to obtain tabulation outliers for both historic and current cell ratio tests in the services sectors portion of the Economic Census (Sigman, 2005). For current cell comparisons, the HB edit performs the following series of transformations on the **original** distribution of current cell ratios prior to outlier identification:

- **Centering transformation**

$$s_i = \begin{cases} \frac{\hat{R}_i}{m} - 1 & \hat{R}_i \geq m \\ 1 - \frac{m}{\hat{R}_i} & 0 < \hat{R}_i < m \end{cases}$$

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

- **Magnitude transformation**

$$E_i = s_i \{MAX(\hat{Y}_{C,i}, \hat{\hat{R}}\hat{X}_{C,i})^U$$

where $\hat{\hat{R}}$ is an industry-average or median current cell ratio (I use $\hat{\hat{R}} = m$) and $0 \leq U \leq 1$. The industry-average ratio ensures that both of the ratio items are converted to the same units of measure.

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 Sigman (2005) and Banim (2000), I 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($E_m - E_{Q1}$,|A E_m|) and d_{Q3} = MAX($E_{Q3} - E_m$,|A E_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. I use A =0.05, as recommended in the original paper (Hidiroglou and Berthelot, 1986) and examine rules determined with C=10 and C=20.

2.2. Multivariate Outlier Detection Methods

With multivariate outlier detection, I consider the variables jointly to identify estimation cells that have outliers in **several** different variables. The Mahalanobis distance statistic is the classical method used to identify outliers in R^p (p =the number of covariates) from a randomly sampled dataset ($\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$), where $\boldsymbol{x}_i' = (x_{i1},,x_{ip})$, assumed drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, estimated respectively by T(**X**) and C(**X**). For each observation i, the Mahalanobis distance is computed as $MD_i = (x_i - T(\mathbf{X}))C(\mathbf{X})(x_i - T(\mathbf{X}))'$. Outlying observations are identified by comparing each computed MD_i to a χ_p^2 critical value. The assumption of multivariate normality with economic macro-data is somewhat questionable, although multivariate lognormality is common.

The classical Mahalanobis distance is prone to masking effects because of the weak resistance from the parametric estimators for T(**X**) and C(**X**). Ghosh-Dastidar and Schafer (2006) describe using maximum-likelihood estimation (M-estimation) for T(**X**) and C(**X**). A drawback of M-estimation is that one must estimate the unknown fraction of contaminated data and the variance inflation factor for contaminated data in advance. Moreover, the M-estimation methods have a low breakdown point (at most $1/p$), so that they become considerably less outlier-resistant as the number of covariates increases. Rouseeuw and Zomeren (1990) proposed the use of a minimum volume ellipsoid (MVE) measure 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. Rousseeuw and Van Driessen (1999) developed an alternative algorithm that searches for the mimimum covariance determinant (MCD) that can be obtained from a subset of half the data then develops robust estimates of T(**X**) and C(**X**) with approximately 50% breakdown points that have better asymptotic properties than the MVE estimates. Both the MVE and MCD methods should have low incidence of masking because of their high breakdown points.

3. The Annual Capital Expenditures Survey

ACES 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. 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, but I consider only roster data items.

The ACES data undergo extensive micro-review before macro-level analysis. However, editing and imputation of capital expenditures micro-data present some unique challenges. Standard economic data editing techniques such as ratio editing are generally not applicable at the individual company level. Capital expenditures within the same company are generally characterized by low year-to-year correlation: e.g.,, a company that purchases new computers one year is unlikely to invest much in new equipment expenditures in the following year. This renders historic cell comparisons ineffective – at both the micro- and macro-levels. Current cell ratio comparisons can also be quite misleading, since capital expenditures are often poorly correlated with available auxiliary data such as payroll or receipts (especially for small companies) and the expenditure items within a given company can be poorly correlated, since structural expenditures and equipment expenditures are driven by different needs. Consequently, ACES micro-editing procedures tend to focus on maintaining additivity between reported totals and associated details. During a separate phase of the micro-level review, analysts compute a "robust" Mahalanobis distance using paired values structures and equipment; the production program iterates twice, first to eliminate potential outliers from $T(X)$ and $C(X)$ calculations by comparing the MD_i statistic to a chi-squared critical value at the 10-percent significance level, then to compute the $T(X)$ and $C(X)$ from the reduced data set. I refer to this procedure hereafter as the "production-MD" method. Outlying sample units are down-weighted.

ACES publishes expansion estimates of totals. These expansion estimates are Horvitz-Thompson estimates adjusted for unit non-response via a weight adjustment procedure. ACES also publishes year-to-year change estimates for key statistics.

4. Evaluation Study

4.1. Preliminary Analysis and Classification

This analysis uses data from the 2002 and 2003 ACES data. The input data are the unit non-response adjusted weighted reported data, with edited values substituted for missing reported data items. The final data are the unit non-response adjusted weighted edited/corrected data. I calculated input data and final data estimates by summing the non-response adjusted weighted data (with outlier corrections) using the company-reported industry as classification variable. The 2002 ACES data set contains 137 industry-level records; the 2003 ACES data set contains 148 industry -evel records. The **input data estimates** are my analysis variables, and the final data estimates are my gold standard" estimates (for evaluation). To begin, I 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.

Outlier (O). The percentage difference between the input and final data estimates in this cell was greater than the 99^{th} percentile of the distribution for the item in the collection period.
Potential outlier (P). The percentage difference between the input and final data estimates in this cell was between the 95^{th} and 99^{th} percentiles of the distribution for the item in the collection period.
Not an outlier (N) The item in this cell was not flagged as an "O" or "P."

An obvious limitation of this classification procedure is the subjective selection of the 95^{th} percentile as an outlier "cut-off." This decision was data-based and is not meant as a recommendation for other data sets.

A **conservative** evaluation classifies **input** data estimates as follows:

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.

An **anti-conservative** evaluation classifies **input** data estimates as follows:

Bivariate: a ratio pair of estimates is an outlier if either the numerator or the denominator is flagged as an **outlier (O) or a potential outlier (P)** 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) or a potential outlier (P)** and is not considered an outlier otherwise.

Counts of estimate classifications for the bivariate and multivariate comparisons are available upon request.

Once each record was classified as an outlier, I applied the bivariate and multivariate outlier detection methods to all **input** data estimates and used the (bivariate or multivariate) outlier classifications to compute the evaluation statistics described in Section 4.2.

Neither the robust regression models nor the resistant fences methods can effectively identify outliers when the estimates being compared have weak statistical association. To determine viable ratio tests for current cell ratios, I calculated Pearson correlation coefficients for each set of ratio tests proposed by our subject-matter experts using the **final** data industry-level estimates. With these data sets, only **two** ratio tests were consistently highly correlated in both data sets: New Structures to Structures and New Equipment to Equipment. Consequently, I restricted my analysis of the robust regression and resistant fences methods to comparisons of New Structures with Structures and New Equipment with Equipment. Since the HB edit is designed to develop flexible limits in the presence of poorly-related or highly volatile ratios, I increased its evaluation set to include comparisons of Structures to Total, New Structures to Used Structures, and Equipment to Total.

4.2. Evaluation Methodology

To evaluate the bivariate methods, I calculated the following statistics for each individual test:

Type I error rate. The proportion of **non-outlier** estimates that are flagged as outliers by a given procedure.
Type II error rate. The proportion of **outlier** estimates that are not flagged as outliers by a given procedure.
Hit rate. The proportion of flagged estimates that are outliers.

The Type I and Type II error rates consider each bivariate comparison to be a hypothesis test, where the null hypothesis is that neither of the tested estimates is an outlier (Thompson and Sigman, 1999). Type I error rates are computed with respect to **non-outlier** observations, and Type II error rates are computed with respect to **outlier** observations. Hit rates (Granquist, 1995) measure the operational efficiency of a given outlier-detection rule. Type I and Type II error rates for individual tests are controlled by modifying test rule parameters. Thompson and Sigman (1999) note that when data items are subjected to more than one bivariate edit or are considered jointly in a multivariate outlier detection procedure, then individual Type I and Type II error rates are a poor measure of the **unidentified** outliers in the completely reviewed data set. A better measure is the all-item Type II error rate, defined as the proportion of **outlier** estimates that are **not** flagged as outliers by any given procedure within a set of outlier detection procedures (with respect to the total set of identified outliers).

Sets of "bad" multivariate observations differ, depending on the outlier-detection procedure. For the joint set of resistant fences and robust regression tests, multivariate classification for a given set of industry estimates consider structures, new structures, equipment, and new equipment (the tested items). For the joint set of HB edit tests and the multivariate techniques, all six estimates are considered, since each is compared in at least one ratio test.

For the multivariate analysis, I calculate Type I and Type II error rates, where 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.

To save space, I present only the anti-conservative results. The complete set of comparisons is available upon request. Throughout the remainder of the paper, I use the abbreviation of STRUCT for Structures and EQUIP for Equipment, and suffix with an _N (new) or _U (used) as applicable.

4.3. Results

4.3.1. Bivariate Comparisons

I began by examining the single test evaluation statistics for the current cell ratio tests from the 2002 and 2003 data sets, respectively (results available upon request). I found that an artifact of my estimate classification procedure was that the Type II error rates increase with the anti-conservative method over the conservative while the Type I error rates may remain constant because many of the "potential" outlying estimates are relatively close to the majority of the distribution.

I found no advantage in using robust regression methods over resistant fences methods for the two highly-correlated ratio tests. In fact, the resistant regression approach is more prone to swamping than the other bivariate methods, with the slight decreases in the Type II error rates being offset by the corresponding increases in the Type I error rates. For both the resistant fences and HB edit applications, there are clear advantages to defining a narrower outlier-detection range in terms of controlling the Type I error rate while decreasing the Type II error rate. With the ACES data, symmetrizing distributions of ratios prior to applying resistant fences rules does not appear to further improve the effectiveness of the ratio edits. Of the remaining methods, the HB edit with $U = 0.3$ and $C = 10$ best balanced the individual Type I and Type II error rates (although the individual Type II error rates are not small), while yielding fairly reasonable hit rates, although it does not outperform the asymmetric resistant fences applications for the high-correlation ratios.

Because the estimates are tested in more than one ratio test, it is necessary to consider all tests **jointly** to assess the overall proportion of outliers left undetected after all ratio edits are examined. Table 1 presents the all-item Type II error rates for each method and data set

Table 1: All-item Type II Error Rates

Outlier Method	Items	2002	2003
Robust Regression	STRUCT, STRUCT_N, EQUIP, EQUIP_N	0.57	0.43
Asymmetric Inner Fences	STRUCT, STRUCT_N, EQUIP, EQUIP_N	0.36	0.37
Asymmetric Outer Fences	STRUCT, STRUCT_N, EQUIP, EQUIP_N	0.42	0.39
Symmetric Inner Fences	STRUCT, STRUCT_N, EQUIP, EQUIP_N	0.38	0.39
Symmetric Outer Fences	STRUCT, STRUCT_N, EQUIP, EQUIP_N	0.47	0.43
HB (U = 0.5, C=10)	All Items but EQUIP_U	0.45	0.36
HB (U = 0.5, C=20)	All Items but EQUIP_U	0.52	0.42
HB (U = 0.3, C=10)	All Items but EQUIP_U	0.42	0.36
HB (U = 0.3, C=20)	All Items but EQUIP_U	0.51	0.44

The all-item Type II error rates for the robust regression, symmetric resistant fences, and HB edit tests with $C = 20$ are considerably higher than those from the asymmetric inner fences and HB edit ($U = 0.3$, C=10) tests. Although the all-item Type II error rates are quite good with the Symmetric Inner Fences data in 2002, these same tests were not at all effective in terms of Type I error rates and do not test as many items as the joint set of HB edit tests. Moreover, the HB edit tests outperform the asymmetric inner fences tests on the 2003 data sets, while comparing more items.

4.3.2. Multivariate Method Comparisons

I considered two different sets of multivariate observations:{STRUCT,EQUIP}and {STRUCT_N, STRUCT_U, EQUIP_N,EQUIP_U}. The first set is used in the ACES **micro level** outlier detection procedures. The other set is intuitively appealing, being comprised of mutually exclusive estimates. Following the suggestion of Franklin, Thomas, and Brodeur (2000), I consider the MVE and MCD applications on log-transformed data (substituting 0.000001 for the undefined log(0)) as well as on the original data. For the MVE and MCE applications, I drop observations that consist of entirely-zero estimates (due to software limitations). I apply the ACES "production-MD" to the full original data only; neither the log transformation nor the dropped entirely-zero estimates had any effect on the calculated statistics.

The first set of comparisons on {STRUCT,EQUIP} were not promising, with each method characterized by an unacceptably high Type II error rate. Table 2 presents the Type I and Type II error rates for the second set of Mahalanobis distance comparisons.

Table 2: Type I and Type II Error Rates {STRUCT_N, STRUCT_U, EQUIP_N,EQUIP_U}

Outlier Method	2002		2004	
	Type I	Type II	Type I	Type II
Production-MD	0.00	0.88	0.00	0.89
MCD (original)	0.45	0.47	0.36	0.28
MVE (original)	0.39	0.48	0.33	0.30
MCD (log)	0.04	0.59	0.03	0.65
MVE (log)	0.03	0.62	0.05	0.56

These results for the log-transformed MCD and MVE Mahalanobis distances are far less discouraging, since most of the flagged cases are indeed outliers. Of the two resistant methods, the MVE results have a marginally better Type I Error/Type II Error balance. Without using the log-transformation prior to multivariate outlier identification, all "flavors" of Mahalanobis distance are very prone to swamping. This is not unexpected, given the inherent assumption of multivariate normality (a weaker assumption for the robust methods).

4.3.3. Direct Comparison of Multivariate and Bivariate Methods

The Type I and Type II error analyses presented in Sections 4.3.1. and 4.3.2. provide measures of overall effectiveness for each of the analysis techniques, but do not compare their respective performance. Table 3 does this. When the bivariate and multivariate outcomes differ, the combination of HB edits flags **more** outliers than the multivariate methods. Consequently, the all-item Type II error rates for the complete set of HB edits are lower than the corresponding MVE error rates (for the 2002 data, the anticonservative Type II rate is 0.47 (HB) and 0.62 (MVE); for the 2003 data, the anticonservative Type II rate is 0.36 (HB) and 0.56 (MVE)). Unfortunately, the jointly considered HB edits also have a higher Type I error rate than the MVE error rates, which has analyst workload implications. Consequently, it would be difficult to recommend using this approach as the **sole** method of outlier detection. Nonetheless, the multivariate approach could be a promising first approach, which could be reinforced by examining a few of the current cell ratio tests (with the HB edit) after investigating the initial sets of multivariate-method flagged estimates.

Table 3: Comparison of Multivariate Outlier Procedure and (Joint) Bivariate Procedure

Survey Year	Outcome	Total Cases	Outlier Procedure Outcome	HB Edit		MVE	
				True Outlier	True Non-Outlier	True Outlier	True Non-Outlier
2002	Same	116	Not Outlier	27	67	27	67
			Outlier	20	2	20	2
	Different	32	Not Outlier	8	0	18	6
			Outlier	18	6	8	0
2003	Same	110	Not Outlier	17	69	17	69
			Outlier	22	2	22	2
	Different	22	Not Outlier	2	2	13	5
			Outlier	13	5	2	2

5. Discussion

Our work would be simpler if one could apply a "one method fits all" approach to macro editing. That approach might even work if we could guarantee that our estimates – or pairs of estimates - have the same (or similar) distributions from sample to sample – or even within sample. It is unlikely that the latter condition holds with the survey examined here, but may be quite realistic for another periodic survey.

In this paper, I examined a set of bivariate and multivariate outlier detection methods on economic data that have inconsistent statistical association between items from one collection period to another. With these data, one simply cannot "hard code" any limits: all critical values must be determined by the data set at hand. With the ACES data, dynamically standardizing the comparisons, then dynamically computing outlier limits did quite a bit to ameliorate the effect of distributional differences. However, each 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.

For ACES, I believe that a hybrid approach to macro-level editing could be very effective: multivariate MVE outlier identification, followed by a current cell ratio comparison. Along the same lines, an automatic procedure that flags bivariate pairs via the HB edit tests, then unduplicates records could be equally more effective, since more items are scrutinized. For multivariate analysis, my best results were obtained by log-transforming the estimates, then using the minimum volume ellipsoid (MVE) Mahalanobis distance measure to identify outliers. In terms of bivariate analysis, my best results are obtained with some form of the HB edit.

This paper examines one set of economic data and considers only two separate collections from this program. To extrapolate my recommendations to other data sets would be foolish, especially considering the relatively atypical nature of the economic data and the subjectivity of the outlier classification methods. These results need to be validated on other economic data sets – perhaps even a more typical periodic business survey or via a well-constructed simulation study.

Acknowledgements

The author thanks Broderick Oliver, Jeffrey Dalzell, 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

Bienias, J.L., Bell, T., Caldwell, W., Garrett, V., Hall, I., Hogan, H., Hundertmakr, D., Juzwiak, J., Knowlton, W. Lassman, D., Scheleur, S. , Stachurski, D., Wright, G. (1994). Graphical Review of Economic Data. Internal Census Bureau memorandum, document ESMD-9406, available from the Office of Statistical Research and Methods for Economic Programs.

Cornett, E., McLaughlin, J.F., Hogue, C.R., and Owens, S.D.(2006). A Comparison of Two Ratio Edit Methods for the Annual Survey of Government Finances. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, to appear.

Fescina, R.M., Jennings, A.S., Wroblewski, M.J. (2004). Automated Production of Foreign Trade Data Edit Parameters Using Resistant Fences. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp .3492-3497.

Franklin, S., Thomas, S., and Brodeur , M. (2000). Robust Multivariate Outlier Detection Using Mahalonobis Distance and Modified Stahel-Donoho Estimators. *Proceedings of the Second International Conference on Establishment Surveys*, pp. 697 – 716.

Ghosh-Dastidar, B. and Schafer, J.L (2006). Outlier Detection and Editing Procedures for Continuous Multivariate Data. *Journal of Official Statistics*, **22**, pp. 487 – 506.

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

Hoaglin, D.C; Mosteller, F; and Tukey J.W. (Eds.) (1983). Understanding Robust and Exploratory Data Analysis. NY: Wiley.

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

Rousseeuw, P.J. (1984). Least Median of Squares Regression. Journal of the American Statistical Association, **79**, pp .871-880.

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.

Rousseuw, P.J. and Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**, pp. 212-223.

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.

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.

Thompson, K.J., Sausman, K., Walkup, M., Dahl, S., King, C.S., and Adeshiyan, S.A. (2001). Developing Ratio Edits and Imputation Parameters for the Services Sectors Censuses Plain Vanilla Ratio Edit Module Test. Internal Census Bureau memorandum, document ESM-0101, available from the Office of Statistical Research and Methods for Economic Programs.