RESEARCH REPORT SERIES
(Statistics #2012-10)

Score Functions for Selective Editing of the US Census Bureau Trade Data

María García
Emily Bartha

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: August 20, 2012

Score functions for selective editing of the US Census Bureau Trade Data[1]

María García and Emily Bartha

ABSTRACT

The Foreign Trade Division at the US Census Bureau is responsible for the production and publication of monthly import and export statistics. These data are not survey based, but collected electronically through an online internet filing system upon arrival or departure of merchandise goods. The validity of these data is checked and verified at every step of collection, processing, and tabulation. In this report we present score functions for selective editing of these data. The scores have two components: a measure of how suspicious an incoming record is and a measure of the impact changes in the record may have on publication totals within a particular set of commodity groupings. We also present results of a feasibility study on the application of selective editing to the checking and correction phase of foreign trade exports data processing.

I. Background on Foreign Trade Data Processing

The Foreign Trade Division (FTD) at the Census Bureau processes monthly import and export transactions for the shipment of merchandise between the United States and its international trading partners and publishes the official international trade statistics for the country. Foreign trade transactions are filed via an online internet data collection system, mostly through the U.S. Customs and Border Protection. The collection of these data is unusual at the Census Bureau because they are filed upon arrival or departure of merchandise goods and are not based on surveys or censuses that are sent to respondents soliciting responses.

Data items collected include commodity, country of origin or destination, port of arrival or dispatch, value, quantity, and shipping weight. Data processing begins with extensive micro-editing using the division's automated edit and imputation system that uses a parameter file called the Edit Master. The Edit Master verifies that numeric data fall within the prescribed ranges and that the ratios of highly correlated items fall within prescribed commodity bounds. Records that do not pass the edits are automatically imputed. However, imputation may not be successful for a small portion of the edit failing records. Records for which imputation is not successful are marked as "rejects" and distributed by commodity and sent to subject matter experts for manual review. The analysts use their commodity expertise to manually adjust rejected records. They may also call back filers in an attempt to correct erroneous data. The commodity experts review a large number of records under tight time-constraints before the publication of monthly statistics deadline. Due to the time and resource constraints, the division has an ongoing effort to improve the current procedures while preserving (or improving) data quality. To this aim, we are investigating the feasibility of a selective editing application to these data.

[1] The authors thank Ryan Fescina and William Yancey for their review and helpful comments. This report is released to inform parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily of the U. S. Census Bureau.

In this section we provided background on foreign trade data editing procedures. In Section II we present background on selective editing. In Section III we present the score functions and in Section IV we present results of the feasibility study. We close with a short summary in Section V.

II. Background on Selective Editing

The manual review and follow-up of suspicious units consumes a large amount of data editing resources. In selective editing, this cost is reduced by concentrating the review effort on erroneous units with a large potential impact on publication figures. A score function is used to rank records; records with a score higher than a preset cut-off value are prioritized for manual review according to their score. All other records are either not edited or edited using an automated system. The overall objective is to spend manual review resources on suspicious records that may have a significant impact on the estimates without affecting overall data quality. Research has shown selective editing methods potential for reducing editing costs without affecting the quality of the final publication estimates. Greenberg and Petkunas (1986) report research for an economic survey in which as much as five percent of the erroneous units were responsible for 90 percent of the published estimates. They conclude that a thorough follow-up of all erroneous units had little effect on the final publication totals. Lindell (1997) reports on a study in which the highest ranked 20 percent of the erroneous records contribute to 90 percent of the total adjustment. Granquist and Kovar (1997) showed that selective editing can produce savings of 50 percent or more of the total editing cost while having a small impact on the final publication. Latouche and Berthelot (1994) developed score functions for an annual retail trade survey and Lawrence and McDavitt (1994) presented a score function for a quarterly average weekly earnings survey. Jäder and Norberg (2005) developed a score function for the Swedish foreign trade survey.

Garcia et al. (2007) developed score functions for prioritizing manual review of records rejected ("rejects") by the trade statistics editing system. Their research focused on prioritizing manual review of suspicious records for which the Edit Master imputation procedures failed to find acceptable imputes. In this research we investigate the feasibility of using the methodology earlier in the editing process. We also present alternative ways to calculate the suspicion term in the score function. This editing strategy uses the score functions prior to the Edit Master. The procedure will streamline records into a two-tiered system. Records are ranked according to their scores; records with scores higher than a preset cut-off value are marked for manual review. All others records are accepted as handled by the automatic editing system. This process will allow a more efficient target of records for review and identify highly influential errors requiring manual intervention earlier in the editing process.

III. Score functions for selective editing of trade data

The current Edit Master editing processing includes ratio edits involving items Value, Quantity, and Shipping Weight denoted by variables *V*, *Q*, and *SW*, respectively. For any given shipment, the ratio of value to quantity, $p = V / Q$, denotes the unit price of the shipped merchandise. For this study we focus on the unit price ratios only.

Hidiroglou-Berthelot method

The Hidiroglou-Berthelot edit (Hidiriglou and Berhelot, 1986) detects outlying ratios in periodic data. For the trade data it is not possible to use this method as described by Hidiroglou and Berthelot; the method must be adapted for application to current ratios. For every record *i*, the Hidiroglou-Berthelot edit as applied to this data begins with the current month unit price ratio $p_i = V_i / Q_i$ and the median of unit prices p_{q_2}. The unit price ratios p_i are transformed to ensure possible outliers are identified at both ends of the distributions,

$$S_i = \begin{cases} p_i / p_{q_2} - 1 \; if \;\; p_i > p_{q_2} \\ 1 - p_{q_2} / p_i \;\; if \;\; p_i < p_{q_2} \\ 0 \; if \;\; p_i = p_{q_2} \end{cases}$$

Note that this transformation centers the distribution of ratios about zero; it does not provide a symmetric distribution of the unit price ratios.

With business data, we wish to ensure we are tracking errors associated with large units that affect many statistics the most. Hidiroglou and Berthelot (1986) suggest applying another transformation to exercise control over the influence of the magnitude of the data. This transformation compares current cycle data to previous cycle data to ensure more importance is placed on a small deviation within a large unit as opposed to large deviations within small units. We recall that this method was developed for periodic data, where previous and current data are used to identify suspicious units. This is not the case with the trade data. In these data, for most commodities previous month data may not be available or comparable to current month data. Companies may have *m* number of shipments the current month and $n \neq m$ (or no shipments) the previous month. The magnitude transformation must be adapted to using only current cycle data.

When using only current month unit prices, the median of unit prices and reported data can be used to estimate an anticipated value for current month data (see Garcia et al., 2008). Since we are dealing with current month unit price ratios ($p_i = V_i / Q_i$) we use $p_{q_2} * Q_i$ as the best possible anticipated value of V_i , with p_{q_2} denoting the median of current month unit prices as defined above. The size transformation adapted for current month unit price ratios is,

$$E_i = S_i * \{\max(V_i, p_{q_2} * Q_i)\}^u, \text{where } 0 \le u \le 1.$$

In this application we use the square root in the maximization part of the size transformation (i.e. $u = 0.5$).

Let E_{q_1}, E_{q_2} and E_{q_3} be the first quartile, the median and the third quartile of the transformed unit price ratios E_i . We calculate a measure of the deviation of the first and third quartile of the transformed unit price ratios from the median as,

$$d_{q_1} = \max(E_{q_2} - E_{q_1}, | a * E_{q_2} |)$$
$$d_{q_3} = \max(E_{q_3} - E_{q_2}, | a * E_{q_2} |)$$

We assign to every observation a score that is the ratio of the displacement of the transformed unit prices from the median and the appropriate distance from the median as measured by d_{q_1} and d_{q_3},

$$Ratio_i = \begin{cases} (E_{q_2} - E_i)/d_{q_1} & if \;\; E_i < E_{q_2} \\ (E_i - E_{q_2})/d_{q_3} & if \;\; E_i > E_{q_2} \end{cases}$$

The $|a * E_{q_2}|$ in the calculation of the distances ensures that d_{q_1} and d_{q_3} are not too small for observations clustered about the median. We used the value suggested by Hidiroglou and Berthelot ($a = 0.05$) as it has worked well in our application.

Effect of errors on publication totals

The score function *Ratio* can be seen as a measure of how suspicious a record is. We would like to also consider a measure of the relative effect errors in a record have on publication totals. The measure we used is adapted from the *Diff* function described by Latouche and Berthelot (1992). In their study, all variables are used in the calculation of the effect on totals. We consider only the effect of changes in *V*. The effect is calculated as the absolute difference between the current months's reported values and an anticipated value for the current month data. Latouche and Berthelot used final values from the previous cycle as the best anticipated value. As this is not possible with the trade data, the effect measure must be adapted to using only current month data. We proceed as in the definition of E_i letting the product of median unit price ratios and quantity ($p_{q_2} * Q_i$) be an anticipated value for *V*. The *Diff* score as adapted for our current month ratios is,

$$Diff_i = \frac{|V_i - p_{q_2} * Q_i|}{Total(V)}$$

The estimated total $Total(V)$ is calculated using reported data for the current month. For commodities with large deviations in the monthly totals we may consider using annual cumulative totals.

Combine suspicion and effect

In selective editing we may consider a composite global score function that includes measures for both suspicion and effect on publication totals. The score is based on how suspicious a record is (*Ratio*) and the impact the suspicious records has on publication totals (*Diff*),

$RatioDiff_i = Ratio_i * Diff_i$.

Jäder and Norberg (2005) proposed a score function that is somewhat similar to *RatioDiff*. However, their measure of suspicion uses the interquartile range rather than d_{q_1} and d_{q_3}. In addition their score includes a term assigning a measure of suspicion to errors in variable *V*. For these data, subject matter experts give higher importance to the variable representing the value of shipments (*V*) over the variables representing quantity and shipping weight. As a consequence, given an error in the unit price ratio edit $p = V / Q$, analysts attempt to change *Q* rather than *V*. Thus we decided to not include a measure of suspicion of errors in variable *V* over errors in the variable *Q* into the score function.

Variations for *Ratio*

The first transformation in the Hidiroglou-Berthelot method is an attempt to account for outliers at both ends of the distribution of unit price ratios. The log transformation can also be used to make the distribution of ratios more symmetric as in the quartile method. We begin by applying this transformation to the data, calculating the quartiles of unit price ratios using the log transformed unit price ratios instead of the E_i's described above before computing d_{q_1}, d_{q_3} and *Ratio*. In this case,

$$d_{q_1} = \log(p_{q_2}) - \log(p_{q_1})$$
$$d_{q_3} = \log(p_{q_3}) - \log(p_{q_2})$$

The suspicion term is computed as,

$$Ratio1_i = \begin{cases} (\log(p_{q_2}) - \log(p_i)) / d_{q_1} \ \ if \ \ \log(p_i) < \log(p_{q_2}) \\ (\log(p_i) - \log(p_{q_2})) / d_{q_3} \ \ if \ \ \log(p_i) > \log(p_{q_2}) \end{cases}$$

As in the Hidiroglou-Berthelot method there may be commodities for which the median of unit prices p_{q_2} is too close to either the upper and/or lower quartile. In this case the denominator in the distances d_{q_1} and/or d_{q_3} is close to zero and is replaced by a small constant or a fraction of the median.

IV. Feasibility Study

In this report we described a selective editing strategy for identifying highly suspicious records in the Census Bureau trade data. We use score functions for assigning measures for how suspicious a record is and the effect errors in the suspicious record have on publication totals. This approach represents a departure from a previous study in which only rejected records were assigned scores for guiding analyst's review (see Garcia et al., 2008). A selective editing strategy allows the editing process to first assign a score to incoming records; records with a score higher than a pre-set cut-off value are marked for manual review. The objective is to spend the bulk of manual review resources on these records ranked as highly influential according to their scores.

Test Data

For the feasibility study, we used a data file containing a subset of the 2004 exports trade data records consisting of four consecutive months of archived raw and edited (final) data and the Foreign Trade Data division Edit Master parameter file. The Edit Master has the necessary information for how to compute the unit price ratios. The records file has data for several items including value of shipments, quantity of shipments, shipping weight, country of destination, mode of transport, and port of dispatch among others.

The number of records within a commodity is important. The score functions include a term based on quartiles of unit price ratios within each commodity. If the number of records is too low then outliers in the distribution of ratios may be included in the computation of the quartiles. As in our previous study (Garcia et al., 2008; also see Fescina et al., 2004), we decided to include only commodities having at least 30 records in the data file. Our final testing data file extract has records for four consecutive months with commodities having at least 30 records within the month. This test file has only a small proportion of the exports data file, however it is suitable for a feasibility study.

Selective Editing Program

We had previously written code to prioritize manual review of records labeled as rejects by the editing system. The assumptions and procedures are different when considering the full data set as opposed to only rejects. Application to the full data set requires re-writing the selective editing legacy routines. We wrote a new SAS macro implementing selective editing using the different score functions described in Section 3. Starting with the Edit Master parameter file and trade records files as input, the program applies selective editing to produce an output file with a priority ranked list of records according to measures of how suspicious a record is and the potential impact it has on publication totals. The program first extracts the commodities that are usable for selective editing according to the instruction described above. It then merges the usable data records to the Edit Master. Once these two files are merged, the program uses the Edit Master's instructions to compute the unit price ratios. With the unit price ratios available, the next programming step is to calculate the medians and quartiles of unit price ratios by commodity. The statistics are then used to compute the measures of impact and suspicion needed for assigning scores. The measure of impact is computed within the main SAS macro; there is a separate macro for computing each of the different measures of suspicion. The measures of suspicion and impact are then combined into a global score for the record. Once each record had been assigned a score, the program produces an output file using SAS PROC RANK. This final procedure assigns to each record a priority ranking according to its score.

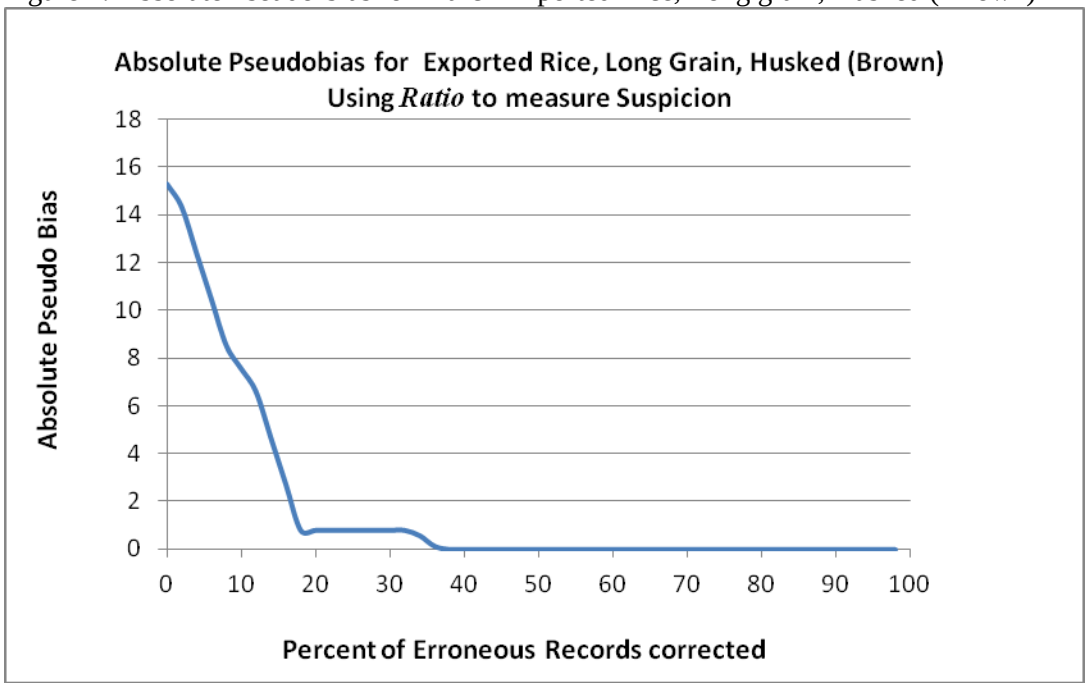
Results of the feasibility study

We study the feasibility of applying the selective editing methodology and associated SAS program with the test data file described above. Latouche and Berthelot (1992) defined a measure to estimate the bias due to errors in the data, called the absolute pseudo-bias, to measure the effectiveness of their score functions. The absolute pseudo-bias, estimates the relative discrepancy between the final publication total T_F, and T_{SE}, the estimated total obtained after processing for selective editing as $|T_{SE} - T_F| / T_F$. In our feasibility study, the

selective editing estimated total T_{SE} is simulated by replacing raw values in records with a score larger than a certain percentage cut-off value with the final data while keeping raw values for records with a score lower than the chosen percentage cut-off value.

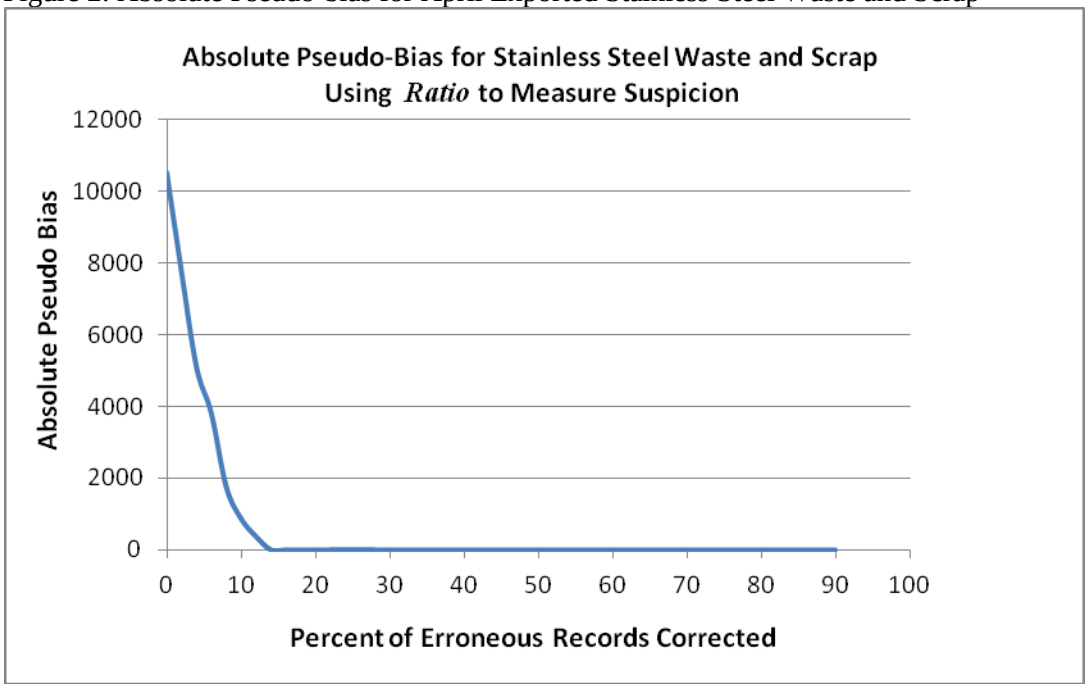
We graphed the *abs*olute pseudo-bias versus the percentage of records corrected for several commodities. On the horizontal axis the records are ordered from the most to the least influential according to their scores before calculating the percentage of records corrected. Figure 1 displays the absolute pseudo-bias for the variable quantity (*Q*) of exported "Rice, long grain, husked (brown)" for the month of March using the score function *RatioDiff* with *Ratio* computed using the Hidiroglou-Berthelot method. The graph illustrates how the absolute pseudo-bias rapidly decreases as the percentage of records flagged for review increases, and review of more than 36 percent of the records with the highest scores does not affect the final estimate. Technically we could stop reviewing records at the 36 percent level of review when the effect of changes on the absolute pseudo-bias approaches zero and the estimated total approaches the final publication total.

Figure 1: Absolute Pseudo-bias for March Exported Rice, Long grain, Husked (Brown)



We observe a similar pattern for the commodity "Stainless steel waste and scrap", with the pseudo-bias decreasing as the number of records corrected increases. For this commodity, Figure 2 shows the absolute pseudo-bias approaching zero as the percent of records corrected hits the 14 percent mark.

Figure 2: Absolute Pseudo-bias for April Exported Stainless Steel Waste and Scrap



There is a caveat in this analysis: there are too many commodities in these data to effectively look at the behavior of the absolute pseudo-bias by commodity. However, it is possible to measure the absolute pseudo-bias at higher levels of aggregation. In a selective editing application, records that are highly suspicious according to their scores are marked for closer scrutiny. Currently, in the foreign trade data review process, erroneous records are classified by commodities and sent to the analysts by commodity groupings called Sections. We thus decided to look at the absolute pseudo-bias by Section. Figures 3 and 4 display the plots of the absolute pseudo-bias for the grouping of commodities (section) "Foods" using *Ratio1* and *Ratio* to measure suspicion respectively. In both graphs we can see the absolute pseudo-bias rapidly decreasing as the number of corrections increases. For this section, measuring suspicion using *Ratio* seems better than measuring suspicion using *Ratio1*; the slope in Figure 4 indicates a fastest decrease of the absolute pseudo-bias as the percentage of records corrected increases.

Figure 3: Absolute Pseudo-bias for April, Section 1, Exported Foods, Using *Ratio1*

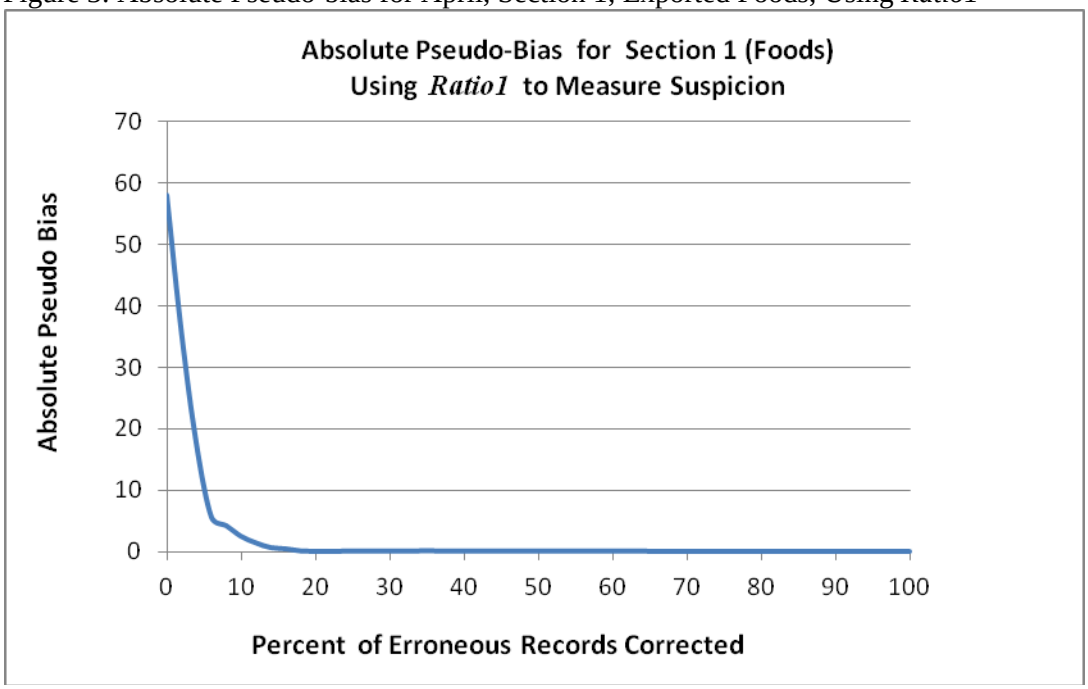
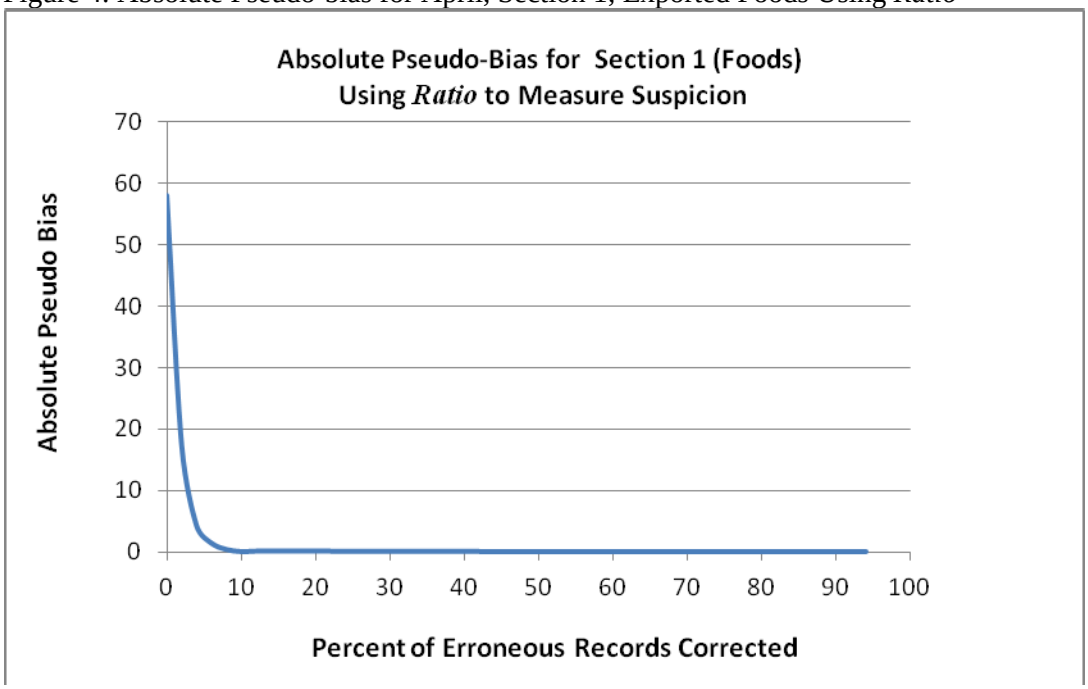


Figure 4: Absolute Pseudo-bias for April, Section 1, Exported Foods Using *Ratio*



Although we cannot display plots for all the commodities within the test data set (there are too many), results showed that it is feasible to apply these score functions as part of the overall foreign trade data editing processing. However this is a very large data set, a study with a larger sample data set is needed. The Foreign Trade division and CSRM are planning a follow up study with a larger data set including more recent data. We expect to get better insight by looking at other measures to evaluate the effectiveness of a selective editing application. One could compare the proportion of records that are flagged as highly suspicious with the proportion of records flagged for manual follow-up by the current editing system. The proportion of records that are flagged can be used as a measure of operational efficiency. However, we were not able to accurately calculate these proportions due to our inability to properly match the data files. Our test data set does not have the unique identifiers needed for the matching operations (Rachelle Reeder, private communication.) CSRM and FTD have recently signed a new data sharing agreement; with a new, larger data set available, we would be able to include this measure in an evaluation study.

Illustrating example: High priority in the ranked list of records

For completeness we include two examples (see Garcia et al., 2008) illustrating the concept of a "high priority" record according to a selective editing application. A major concern for a selective editing application to this data would be the assignment of a high ranking to records that subject matter analysts might consider as not having a significant effect on cell estimates. Table 2 displays analysis for commodity code representing "Blocks, Tiles, and Similar Refractory Ceramic Goods of Clay NESOI exported in March 2004". The Selective editing identified a reported record (shaded in yellow) as having a large impact on the aggregated unit price $p=V/Q$. Existing editing procedures may consider this record as having low importance because it bases the impact the record will have on aggregated totals by value (V) alone. Analysts manually correct the record if the value field fails a range edit; using this criterion the record would have low priority during manual review. However, the difference between changing the record and not changing the record has a significant impact on the quantity (Q) field. We measure the impact changes in this record have on the total quantity Q within the commodity using the percentage relative discrepancy between the reported (T_R) and final (T_F) quantity totals,

$$(|T_F - T_R| / T_F) \times 100 = 8595$$

The record has a large impact (8595 percent) on the total quantity for the commodity and thus it is correctly identified by the selective editing methodology as having a very high priority on the ranked list of records.

Table 3 illustrates the opposite situation: identifying for follow up a record that does not have a high impact on the final cell tabulations. For the commodity code representing "Glass mirrors unframed, not vehicle rearview mirrors" there are 128 reported records with 90 records identified as edit failing records; automatic imputation was successful for 87 records. There are three failing records marked as rejects by the automated system. Although one of the records has only a marginal impact on the aggregated cell, all three records may be corrected during manual review so that the aggregated unit price falls between the prescribed

edit bounds (see yellow row). However, selective editing identified only two records as having a high potential effect on the final totals. Changing the value for quantity *Q* only on these two records (see blue row) brings the aggregated unit price *V/Q* within the optimal bounds for this commodity as desired.

Table 2: Blocks, Tiles, and Similar Refractory Ceramic Goods of Clay NESOI

	Total Value (*V*)	Total Quantity (*Q*)	Unit Price (*V/Q*)	Ratio Bounds	
				Lower Bound	Upper Bound
Reported cell total (10 records)	$102,190	7,217	$14.15	90	3000
Reported suspicious record	$3,024	7,144	$0.42	90	3000
Final suspicious record	$3,024	10	$302.40	90	3000
Final cell total (10 records)	$102,190	83	$1,231.20	90	3000

Table 3: Glass Mirrors Unframed, Not Vehicle Rearview Mirror

	Total Value (*V*)	Total Quantity (*Q*)	Unit Price (*V/Q*)	Ratio Bounds	
				Lower Bound	Upper Bound
128 records, 87 records imputed Three rejects	$3,142,622	129,973,502	$0.02	0.25	50
Final cell total All three rejects corrected	$3,142,622	1,230,629	$2.55	0.25	50
Selective Editing cell total Two highest ranked records corrected	$3,142,622	1,804,699	$1.74	0.25	50

V. Summary

We presented research on score functions for selective editing of the Census Bureau foreign trade data along with results of a feasibility study. In traditional selective editing methods, data from previous cycle are used to construct score functions; this is not possible for the trade data. We adapted available score functions to using only current cycle data by computing an estimate of the anticipated value of the variables. The associated computer program assigns a score to every observation and the output is a ranked listing of records. The ranking is based on measures of how suspicious a record is and the potential impact the record has on the final estimates.

REFERENCES

Fescina, R., Jennings A., Wroblewski, M. (2004). "Automated Production of Foreign Trade Data Parameters Using Resistant Fences". Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association.

Garcia, M., Gajcowski, A., and Jennings, A. (2007). "Selective editing strategies for the US Census Bureau foreign trade data." SRD Research Report RRS 2007/20.

Granquist, L. and Kovar, J. (1997). "Editing of Survey Data: How much is enough?" *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (Editors). Wiley, NY.

Greenberg, B. and Petkunas, T. (1986). "An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division," SRD Report RR-86/04, U.S. Census Bureau. Washington, D.C.

Hidiroglou, M., and Berthelot, J. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". Survey Methodology, V. 12, No. 1, 1986.

Jäder, A. and Norberg, A., (2005). "A Selective Editing Method Considering Both Suspicion and Potential Impact Developed and Applied to the Swedish Foreign Trade Statistics," UNECE Work Session on Statistical Data Editing, Ottawa, Canada.

Latouche, M. and Berthelot, J., (1992). "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys." JOS, V.8 No. 3.

Lawrence, D. and McDavitt, C.,(1994). "Significance Editing in the Australian Survey of Average Weekly Earnings. JOS, V.10 No.4.

Lindell, K. (1997). "Impact of Editing on the Salary Statistics for Employees in County Council." UN Economic Commission for Europe, Statistical Data Editing, vol. 2, Geneva, pp 2-7.

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