

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

**Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)**

Topic (iv): Macro-editing

**SELECTIVE EDITING STRATEGIES FOR THE U. S. CENSUS BUREAU FOREIGN TRADE
STATISTICS PROGRAMS**

Supporting Paper

**Prepared by Maria M. Garcia maria.m.garcia@census.gov, Alison Gajcowski, and Andrew
Jennings U.S. Census Bureau, United States¹**

Abstract

The U.S. Census Bureau Foreign Trade Division (FTD) publishes monthly import and export statistics for the shipment of merchandise goods between foreign countries and the United States customs territories and foreign trade zones. The FTD publishes these statistics for over 17,000 import and 10,000 export commodity classifications. These data are not survey based, but collected from forms upon arrival or departure of merchandise goods. Data are edited and checked at every step of collection, processing, and tabulation. However, due to the limitations of the merchandise trade statistics program, monthly publication cells may still be subject to errors. We present score functions to rank edit failing records according to their potential impact on publication totals. In this paper we present four separate score functions and testing with data from the 2004 foreign trade transactions reporting.

I. INTRODUCTION

1. The Foreign Trade Division (FTD) at the U.S. Census Bureau is the official source for the United States merchandise trade statistics. The FTD publishes monthly import and export statistics for the shipment of merchandise goods between the United States and its international trading partners providing a comprehensive enumeration. Transactions are filed via electronic or paper means, mostly through the U.S. Customs and Border Protection (Customs). 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. The data are classified using the Harmonized Commodity Classification System (HS) by assigning a 10-digit code to each commodity. The division receives about 3.4 million import records and 1.8 million export records every month covering 17,000 imports and 10,000 export commodity classifications.

2. These data are edited and checked for errors and quality assurance at every step of collection and processing. The data are subject to extensive micro-editing using the division's automated edit and imputation system that uses a parameter file called the Edit Master (EM). The EM verifies that numeric

¹The authors wish to thank Andreeana Able, April Downs and Sharon Ennis for their contribution on this project. We also wish to thank David Dickerson, Ryan Fescina, Yves Thibaudeau, and William Winkler for their helpful comments on earlier versions of this paper. 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.

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 or rejected based on a value of shipments threshold. However, imputation may not be successful for a small portion of the edit failing records. Records for which imputation failed are marked as “rejects” and require manual resolution. Each month, fewer than 0.5% of the 5 million import/export records are rejects. The analysts use their commodity expertise to manually adjust rejected records. They may also call back filers in an attempt to correct erroneous data.

3. Manual review and follow-up of suspicious units consume a large amount of the data editing resources. In selective editing, this cost is reduced by concentrating the review effort on erroneous units with a large potential impact on the publication totals. In this paper we present research investigating the use of selective editing methodologies for the foreign trade statistics programs. Section II provides background on foreign trade data editing procedures. Section III describes the different methodologies we used and a weighting scheme for our data. In Section IV we provide a discussion along with results of an application to the 2004 export transactions reporting. We close with a short summary in Section V.

II. EDITING THE U.S. CENSUS BUREAU TRADE STATISTICS DATA

4. The U.S. Census Bureau processes import and export transactions and publishes the official international merchandise trade statistics for the United States. Data items collected include commodity, country of origin or destination, port of arrival or dispatch, value, quantity, and weight. Data processing begins with extensive micro-editing. More than 99 percent of rejects are corrected using an automated system. Records for which imputation is not successful are distributed by commodity and sent to subject matter experts for manual review. The commodity experts review a large number of records under tight time-constraints before the publication of monthly statistics deadline. Due to time and resource constraints, the division has an ongoing effort to improve the current procedures while preserving (or improving) data quality. These efforts focus on outlier detection, automatically updating edit parameters, and selective editing.

5. Fescina et al. (2004) use historical data (currently using up to five years of data) to automatically update edit parameters based on the distributions of the data at the individual commodity level. For the value of shipments (V) and quantity (Q) items, they first compute the unit price of an observation, $p=V/Q$; next symmetrize the data using the log transform and compute quartiles of unit prices, q_1 and q_3 , for each commodity’s log-transformed unit price. Their method then identifies a record as being suspicious if the unit price p falls outside the interval $(q_1 - k(q_3 - q_1), q_3 + k(q_3 - q_1))$, where k is a constant.

6. The division also has an ongoing outlier detection project using X-12-ARIMA time series software. The trade statistics program produces monthly estimates for import and export series aggregated to the five digit end-use classification codes. The purpose of the project is to determine the appropriate ARIMA model for each series and identify which series are outliers for the current month.

III. SCORE FUNCTIONS FOR THE CENSUS BUREAU TRADE DATA

7. Records that are labelled suspicious using the method by Fescina et al. (2004) pass through the Edit Master and will either be automatically imputed or sent to the analysts for review. Our aim is not to re-engineer the current editing procedures but to add selective editing strategies for prioritizing manual review of suspicious records for which the Edit Master imputation procedure is not successful. In selective editing, a score function is used to rank edit failing records; records are then prioritized for review according to their score. 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.

8. 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. In both studies, data from previous survey cycles are required for developing score functions and

the corresponding cut-off values. Thompson and Hostetter (2000) developed score functions for the U.S. Census Bureau Annual Survey of Manufactures using both data from previous collection cycle and administrative data when available. Jäder and Norberg (2005) developed a score function including measures of suspicion and potential impact for the Swedish foreign trade survey.

A. Flagging the most important variable

9. In developing score functions for prioritizing manual review of the Census Bureau's trade statistics data we considered scores previously tested at other institutes. Despite the large number of commodity classifications, we needed a score function to prioritize review of some observations at the 10-digit classification level since users may more closely monitor and scrutinize the statistics for particular types of commodities. Latouche and Berthelot (1992) suggest a simple score function that gives prominence to the most important variable. For our trade data, the analysts give higher importance to the variable representing the value of shipments (V) over the variables representing quantity (Q) and shipping weight (SW) of a shipment. Let i denote the observation, k denote the variables in ratio edits (V , Q , and SW), r index reported data, e edited data, cm and pm denote current and previous month respectively, and Z denote the number of items flagged to be imputed. With V marked as the most important variable, the *Flag* score as described by Latouche and Berthelot (1992) is,

$$Flag_i = \sqrt{\max(V_{i,cm}^r, V_{i,pm}^e)} * Z_i.$$

10. For our trade data it is not possible to use this score function as described. 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 or no shipments the previous month. Thus, we need to adapt *Flag* to using only current month data. On his research on outlier detection using the Hidioglou-Berthelot method, Sigman (2005) noted that when using only current month ratios the median of ratios and reported data can be used instead of the previous month data. Let p_2 denote the current month median of unit price ratios $V_{i,cm}^r / Q_{i,cm}^r$. According to Sigman (2005) we could use $p_2 * Q_{i,cm}^r$ instead of $V_{i,pm}^e$ in the maximization part of *Flag*,

$$Flag_i = \sqrt{\max(V_{i,cm}^r, p_2 * Q_{i,cm}^r)} * Z_i,$$

We also considered a composite *Flag* using the variables value (V) and quantity (Q),

$$CFlag_i = \{\sqrt{\max(V_{i,cm}^r, p_2 * Q_{i,cm}^r)} + \sqrt{\max(Q_{i,cm}^r, t_2 * SW_{i,cm}^r)}\} * Z_i,$$

where t_2 is the median of quantity/shipping weight ratios $Q_{i,cm}^r / SW_{i,cm}^r$ computed at the commodity level. Note that items in *CFlag* must be in the same unit of measurement before computing the score. We do this by using imputation factors which have been computed using commodity averages and are available in the foreign trade data Edit Master.

11. As we mentioned before, *Flag* and *CFlag* are to be applied to a set of particular commodities at the 10-digit commodity classification level as chosen by the subject matter experts. Using this type of score at this level of aggregation is not feasible for the whole data.

B. Effect on publication totals

12. Our next score function is adapted from the *Diff* function described by Latouche and Berthelot (1992) which examines the sum effect of changes in variables V and Q by looking at the absolute difference between the current month reported values and the final values from the previous month on the

publication totals. As before, we cannot implement *Diff* using final data from the previous month and it must be adapted to using only current month data. If using only current month data then,

$$Diff_i = \frac{abs(V_{icm}^r - p_2 * Q_{icm}^r)}{Total(V_{cm})} + \frac{abs(Q_{icm}^r - t_2 * SW_{icm}^r)}{Total(Q_{cm})},$$

where $p_2 * Q_{icm}^r$ and $t_2 * SW_{icm}^r$ are estimates of the expected value of shipments and quantity of shipments for unit i respectively. The estimated total for each commodity is calculated using final data for records accepted or automatically imputed by the automated system and reported data for the current month rejects for every observation within the commodity.

13. We may consider measuring only the effect of changes in the variable value of shipments V over the total value of shipments as in,

$$Diff_i^v = abs(V_{i,cm}^r - p_2 * Q_{i,cm}^r) / Total(V_{cm}).$$

In this case *Diff* is similar to the measure of impact within the score function developed by Jäder and Norberg (2005) for the Swedish trade data.

C. Hidiroglou-Berthelot method

14. The Hidiroglou-Berthelot (HB) method uses historical ratios to identify suspicious records (Hidiroglou and Berthelot, 1986). The HB edit as applied to our data begins with the current month unit price ratio, $p_i = V_i / Q_i$, and p_2 , the median of unit prices. The unit price ratios are then transformed using the following transformation,

$$S_i = \begin{cases} p_i / p_2 - 1 & \text{if } p_i \geq p_2 \\ 1 - p_2 / p_i & \text{if } 0 < p_i \leq p_2 \end{cases},$$

15. Hidiroglou and Berthelot suggest applying another transformation that ensures more importance will be placed on a small price deviation within a large unit as oppose to a large deviation within a small unit. For application to the trade data, the HB method must be adapted to using only current month ratios. Sigman (2005) adapted the transformation suggested by Hidiroglou and Berthelot to using current month data only. The transformation as applied to our current month unit price ratios is,

$$E_i = S_i * \{\max(V_i, p_2 * Q_i)\}^u,$$

where $0 \leq u \leq 1$. A value of $u = 0.5$ as in the maximization part of *Flag* seems to work well for our data.

16. We then calculate a measure of the distance of the first and third quartile of the transformed unit price ratios from the median. Let q_1, q_2 , and q_3 be the first quartile, the median and the third quartile of the transformed unit price ratios respectively, calculate

$$d_{q_1} = \max(q_2 - q_1, abs(a * q_2))$$

$$d_{q_3} = \max(q_3 - q_2, abs(a * q_2))$$

17. Then, assign to every observation a score that is a ratio with a factor measuring displacement of unit prices from the median, weighted by the appropriate distance from the median,

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

18. According to Hidioglou and Berthelot the $abs(a * q_2)$ term 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. In our application we used a value of $a = .05$ as suggested by Hidioglou and Berthelot.

19. We also considered a simple variation of *Ratio* using the log transformed unit price ratios as are computed during parameter development (Fescina et al., 2004). In this case we calculate the quartiles of the transformed unit price ratios q_1, q_2 , and q_3 using the log transformed ratios instead of the E_i 's described previously before computing d_{q_1}, d_{q_3} , and *Ratio*.

D. Combine Hidioglou-Berthelot edit and Effect on publication totals

20. We wanted to include another score function based on the score function developed by Jäder and Norberg (2005). They report on successfully implementing a score function for the Swedish trade data as a weighted geometric mean of measures for suspicion, suspicion of errors in V over errors in Q , and potential impact (See Jäder and Norberg (2005) for details), $Score = Suspicion * Suspicion(V \text{ over } Q)^{weight(SuspV)} * Impact^{weight(Imp)}$. For our data, it was decided to not include a measure of suspicion of errors in value of shipments V over errors in quantity of shipment Q into the scores, but we can use the idea of using the geometric mean to calculate an alternative score using the product of the score based on the HB method (*Ratio*) and the effect on publication totals (*Diff*) described above. [Note: We have not tried assigning weights to separate contributions of *Ratio* and *Diff* yet.] For every observation we compute a new score as,

$$RatioDiff_i = Ratio_i * Diff_i.$$

E. Weights

21. Our team worked together with senior analysts assigned to review and resolve edit failures to design a suitable weighting scheme for the data. The analysts reported considering the size of the company in terms of value of shipments when resolving suspicious records: they recognize when value of shipments is large enough to affect tabulations and thus warrants more attention during manual review. Also, analysts' workload is stacked by section (a section is an aggregate of commodities). Using the analysts recommendations, we assign to every observation a weight based on the value of shipments (w_V) by commodity classification and a separate weight for key commodities (w_C) (key commodities as classified by senior analysts) by section. These weights are incorporated into the score functions.

IV. DISCUSSION AND PRELIMINARY RESULTS

22. In this paper we presented score functions for prioritizing manual review of foreign trade data records identified as suspicious by the automated editing system. Selective editing is probably a misleading term: Our aim is different from the traditional selective editing goal of manually reviewing records with a significant impact on tabulations with all other records handled by an automated system. We did not start out to develop a new editing strategy to identify suspicious records, but to prioritize manual review of records that had already been labelled as suspicious. The editing process will be a two-tiered flow system in which the current month suspicious records are identified and fields marked to be changed are automatically imputed. Then, records for which automatic imputation is not successful

would be marked as rejects and assigned a score to provide a ranking for guiding clerical review. The bulk of manual review resources would be spent on the most important records; however we have the mandate that all rejected records are to be reviewed.

A. Application

23. At the beginning of this project we were to focus on prioritizing manual review of records within a small group of commodities requiring more thorough scrutiny. The *Flag* score is applied only at the 10-digit classification level for this small number of products for which end-users need more detailed statistics. Since we are scoring only rejected records, for most commodities there are not enough observations to apply a score function at this level of aggregation. We must group rejected records at higher aggregation levels. Rejects are classified by commodities and sent to the analysts by sections (commodity groupings). On average each analyst must correct over 600 records per month. Since analysts' workload is stacked by section, we calculate the scores to rank records by section. We do not cross-classify the data with other fields (e.g. country of export, mode of transport, port of dispatch) as this will further reduce the number of records for most commodity groupings. The scores *Diff*, *Ratio*, and *RatioDiff* can be computed at different domains or levels of aggregation.

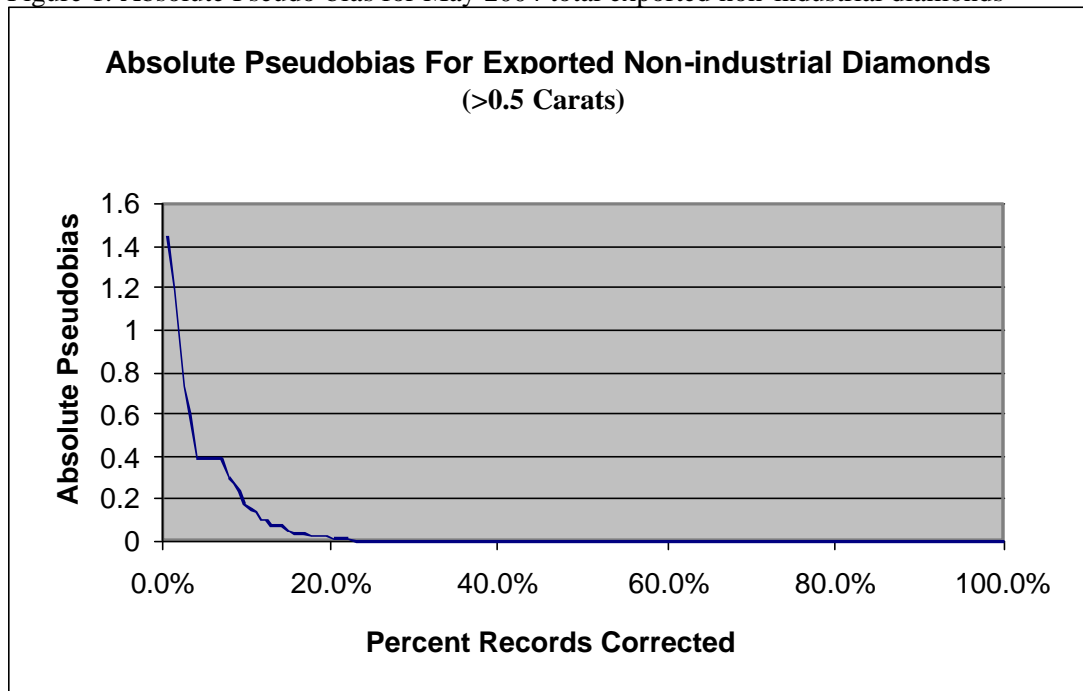
24. We have available archived raw and final data for the 2004 monthly exports transactions reports for products shipments from the United States to foreign countries. The data file has flags for fields marked to be imputed and for records for which automatic imputation was not successful (rejects). The file also contains 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.

25. The medians and quartiles of items and estimated totals in score functions are computed using current month raw data for rejected records and final current month data for all other records. This is possible in our application as items marked to be changed in suspicious records are automatically imputed before labelling records as rejects. (Note: Historical data is used for computing quartiles of data in the measures of suspicion; see Fescina et al., 2004)

26. The purpose of this ongoing study is to investigate the use of a score function for prioritizing manual review of rejected records in the foreign trade statistics program. Since we are mandated to review all rejected records, comparisons between the total value of shipments and quantity of shipments obtained after reviewing records using the different scores are not possible (all records are manually reviewed). A question remains: how do we evaluate the effectiveness of the score functions?

27. In the case of *Flag*, which is used when there are enough records within a particular commodity (say at least 50) for which more detailed statistics are needed we looked at the absolute pseudo-bias. Latouche and Berthelot (1992) define absolute pseudo-bias as $abs(T_E - T_F)/T_F$, where T_F is the final publication total and T_E is the estimated total obtained by replacing raw values in records with a score larger than a certain cut-off value with the final data (keeping raw values for records with a score lower than the cut-off value). Figure 1 displays the absolute pseudo-bias for the variable quantity for shipment of exported non-industrial diamonds in the month of May 2004 (rejected records only) using the score function *Flag*. The pseudo-bias rapidly decreases as the percentage of records marked for clerical review increases, and review of more than 20 percent of the rejected records with the highest scores does not affect the final estimate. Technically we could stop reviewing records at the 20 percent level of review when the effect of changes on the absolute pseudobias approaches 0.

Figure 1. Absolute Pseudo-bias for May 2004 total exported non-industrial diamonds



28. We note that in production all records are reviewed; however comparing the estimated and final totals gives us an idea of how well the score function is tracking the most influential observations. Also, since we are looking at the commodity level, the distributions of scores using *Flag* tend to have a similar shape from month to month (see longer SRD research report, Garcia et al., to appear). In the above example it is possible to expect that if manual review of rejected records proceeds up to top 20% of the ranked records, then the same 20% cut-off value could be used at the next cycle.

29. Our initial approach using the *Flag* score is not suitable for these data at higher commodity groupings. For the other score functions, where records are ranked at higher levels of aggregations, we expect distributions of scores to change in consecutive months due to the magnitude and complexity of the data. In Figures 2 and 3 we showed distributions of the *RatioDiff* scores for the Foods sector for the months of May and June 2004 (only for rejected records, some hidden observations at both ends of the distribution). We zoomed in an area of the graph in which there is an inflection point, the graph stopped “bouncing” around, and the slope of the graph had begun to approach zero. We then fit a trend line to the graph of distribution; the trend line would vary depending upon which area of the graph was zoomed in on. A 95% Confidence Interval was fit around the trend line. An ad-hoc cut-off value could depend upon where the graph of the scores stopped leaving the 95% Confidence Interval. As the trend line would change depending on the area of the graph zoomed in on, estimating a cut-off value this way is rather arbitrary. Since we are mandated to review all rejected records, determining cut-off values for the scores is not an issue.

Figure 2: Trend Lines for the distribution of scores for the May 2004 Food Sector using *RatioDiff*

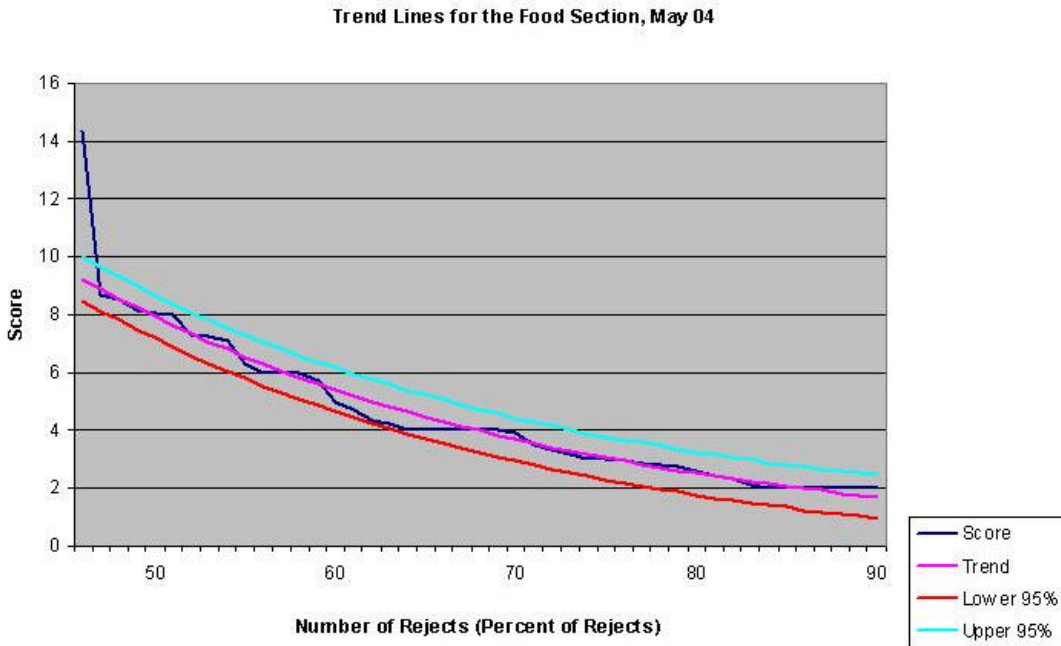
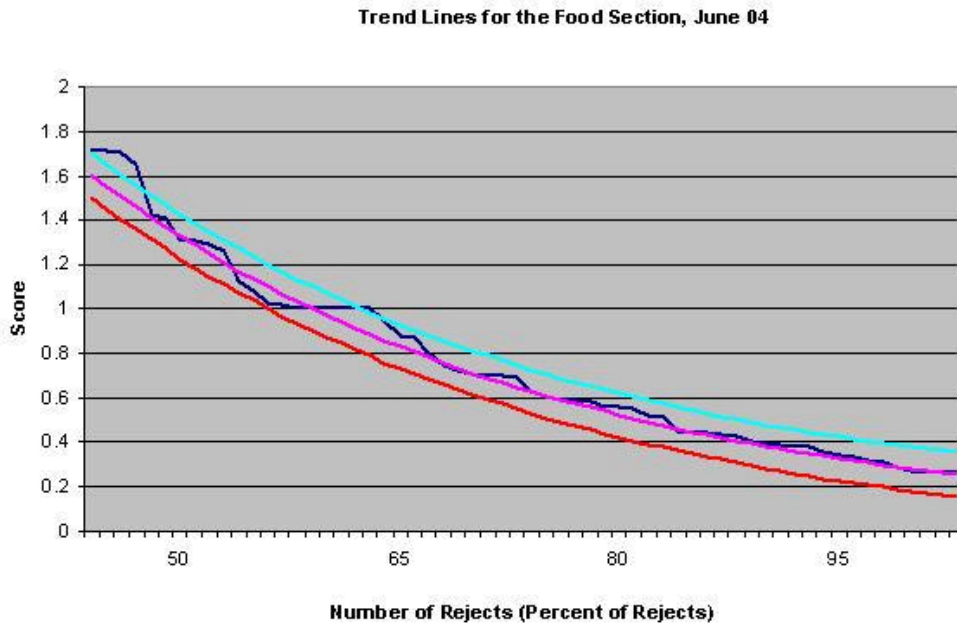


Figure 3: Trend Lines for the distribution of scores for the June 2004 Food Sector using *RatioDiff*



V. SUMMARY

30. In this paper we presented a summary of ongoing research on developing selective editing strategies for the Census Bureau foreign trade statistics program. We presented four separate score functions that can be implemented at different levels of aggregations providing a ranking for guiding clerical review of rejected records. We have not yet provided a recommendation on which score functions to use: determining the best score function when all records must be reviewed requires further study. However, we expect a ranked review process may improve data quality: it provides a ranked order of rejected records that ensures the most resources are spent on the most significant observations. We expect

to begin production testing in the coming months; a detailed evaluation study is to appear in a coming Statistical Research Division report (Garcia et al., 2006)

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., Jennings A., (2006). Results of an evaluation of selective editing strategies for the U.S. Census Bureau foreign trade statistics programs. SRD Research Report Series, to appear.

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

Jader, 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, 2005.

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.

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.

Thompson, K. and Hostetter, S., (2000). Investigation of selective editing procedures for the U.S. Bureau of the Census Economic Programs, Proceedings of the Second International Conference of Establishment Surveys.
