Automated Production of Foreign Trade Data Edit Parameters Using Resistant Fences

Ryan M. Fescina, Andrew S. Jennings, Monica J. Wroblewski, U.S. Census Bureau[1]
Foreign Trade Division, Washington, DC 20233-9100

ABSTRACT

The Foreign Trade Division (FTD) of the U.S. Census Bureau is responsible for publishing the official international merchandise trade statistics for the United States. This includes data for both import and export merchandise trade. Each month, FTD collects and edits approximately 3.4 million import and 1.8 million export records, covering nearly 17,000 import and 10,000 export commodities. These records pass through many edits using multiple sets of parameters. Due to resource constraints and the potential for an increase in the number of parameters, we are investigating ways to automatically create tolerances based on a Symmetrized Resistant Fences method. This paper presents our findings on the implementation of Symmetrized Resistant Fences applied to foreign trade ratio parameters, and the use of historical data to flag questionable parameters or automatically update the parameters as needed.

KEY WORDS: outlier detection, resistant fences, data editing, economic data

1. Introduction

The Foreign Trade Division (FTD) of the U.S. Census Bureau is responsible for publishing the official international merchandise trade statistics for the United States. These statistics are based on all import and export data that are reported to U.S. Customs and Border Protection, whether by paper documents or electronic filing. Federal regulations require most import and export trade activity to be reported. In order to ensure the quality of the published trade statistics, FTD edits the reported data by using a set of edit parameters. The data we process are not survey-based; therefore, we receive and edit approximately 3.4 million import records and 1.8 million export records per month. We classify these imports and exports using the Harmonized Commodity Classification System (HS), which assigns a 10-digit code to each commodity. There are roughly 17,000 import and 10,000 export commodity codes. An import record may be edited against up to 36 different parameters, and an export record may be edited against up to 32 different parameters. These parameters may be different depending on the type of good that is being imported and exported, and thus, we must maintain separate parameters for each import and export commodity code. Our edit parameters are currently updated and maintained manually by subject matter analysts in FTD. Maintaining effective parameters has become increasingly challenging, and an initiative issued by the Secretary of Commerce mandating FTD to accelerate the data release by 7 days will allow even less time for analysts to update the parameters. Consequently, we are investigating new, automated methods to maintain and refresh our parameter files.

In this paper, we propose a new analysis tool for identifying potentially less effective parameters and, in some cases, updating them automatically. This new tool will employ a methodology that will utilize historical data to create reasonable parameters based on the distributions of the data. The tool will be flexible enough that it can be "turned on and off" as necessary, whenever analysts wish to update the parameters, and can either flag questionable parameters for further analyst review, or can automatically update the parameters.

We begin by discussing some background on how we collect and edit foreign trade data. We will then discuss the Resistant Fences methodology that we used to create our new analysis tool, followed by some results of our initial testing of this proposed methodology. Finally, we conclude with a discussion of future analyses.

2. Background on Foreign Trade Data

The import and export data that we collect each month come from two different sources: paper and electronic filing. As the Internet and electronic communication have grown in popularity over the past several years, the number of paper documents we receive each month has steadily decreased. As of September 2003, electronic filing accounted for

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

93.7% of the total value for imports, and 89.3% of the total value for exports. Plans are currently being finalized to make electronic filing of export documents mandatory.

All of the parameters we use to edit the data once we receive it are stored in a file known as an "edit master," which is the primary tool used by our editing programs. For each commodity, the edit master (EM) contains all the appropriate edit parameters that a record for that commodity may be subjected to, as well as codes that tell the program which parameters to edit that record against. Foreign trade data are subjected to two different types of edits: range edits and ratio edits. Range edits check only one value at a time. Examples of range edits for foreign trade data include maximum air shipping weight and maximum vessel shipping weight. If a commodity is shipped either by vessel or air, its shipping weight cannot exceed a certain value. In these cases, only one value (shipping weight) is being edited. We will study the best ways of automating updates to these types of edit parameters at a later date. The parameters that we focused on for this paper are the lower and upper bounds of ratio edits. Ratio edits compare the ratio of two highly correlated items to upper and lower bounds, known as tolerances. An example of a ratio edit for foreign trade data is the unit price or value/quantity (V/Q) edit, which edits the relationship between the value of an item and its reported quantity. Since, for most commodities, value and quantity are highly correlated, their ratio should be contained within a common set of bounds. If a unit price for a particular record falls outside of the tolerances, then that entire record is rejected and must be further analyzed or imputation is performed.

3. Resistant Fences Method

To edit our data more accurately, we need to be able to compare our data against a reliable set of parameters. For certain commodities, the range of acceptable values is defined explicitly in the commodity description. For example, one commodity code is described as 'imports of edge belt sanders, woodworking, new, valued $1000 or over but under $3000 each.' Thus, for this commodity, the lower and upper bounds are fixed at $1000 and $2999.99 respectively. However, the majority of commodities do not have specifically defined lower and upper bounds, and thus, we need a way of determining which values should and should not be accepted for these commodities. We will proceed under the assumption that the majority of foreign trade data are reported correctly. This assumption allows us to define a range in which we expect 'good', or correctly reported, data to fall within. Using the bulk of the data for a commodity, we can produce lower and upper bounds, such that any observation lying outside those bounds is an unusual enough value to be considered an outlier, or in our terminology, a reject.

There are some important issues to keep in mind as we discuss the production of lower and upper bounds for foreign trade unit price ratios. First of all, each commodity has a different number of records each month, so in order to have enough data to produce reasonable limits, we will need to use a large file of historical data. We chose a year's worth of data to increase the number of commodities with a usable amount of records. The unit price values for each commodity have different distributions. It is not correct to assume that the data are distributed identically over all commodities. Lastly, it is impossible for a foreign trade data parameter to be negative. Thus, any lower bounds we create can never be less than zero.

We had to keep all of these considerations in mind when choosing a method to produce our new parameters. We also wanted a method that was not too difficult to program or to explain to the analysts. The resistant fences method satisfies our major requirements, and is the method we will discuss throughout the rest of this paper.

The resistant fences method is an outlier-detection method that uses quartiles of data to determine which data elements are likely in a given distribution of data, and which are unlikely. The basic formula for the lower and upper bounds using resistant fences is:

$$(Q_1 - k*IQR,\ Q_3 + k*IQR)$$

where Q_1 and Q_3 are the first and third quartiles in an ordered distribution of ratios, respectively, IQR is the interquartile range ($Q_3 - Q_1$), and k is a constant, either 1.5, 2, or 3 (Tukey, 1976). The fences that are produced using this formula are called 'inner fences' for k=1.5, 'middle fences' for k=2, and 'outer fences' for k=3. Outliers are defined as any value falling outside of these fences. This method works best for symmetric data, but symmetry is not required. In our preliminary tests, we determined that using unsymmetrized data consistently created new lower bounds that were negative, which are not meaningful in a foreign trade context. Thus, we initially chose to symmetrize our data using a natural log transformation, eliminating the problem of negative lower bounds. Thompson (1999) and Thompson and Sigman (1999) both recommend this power transformation for asymmetric distributions of ratios,

as long as the ratios are positive. After symmetrizing all the data and looking at our results, we will then decide if global symmetrization is appropriate.

4. Creation of Test Parameters Using Resistant Fences

For the purposes of creating new parameters using the resistant fences method, we combine a full year of reported import data from July 2002 to June 2003. It is important to note here that many commodities have two or more different sets of unit price limits in our edit master – one for a specific country whose typical range of data tends to be unique from all other countries, and one for the rest of the world. The unit price range that is not country-specific is known as a 'basket' range. As an example, imports of antidepressants, tranquilizers, and other psychotherapeutic agents from Ireland are subject to lower and upper unit price bounds of $100 and $2500, respectively. By contrast, imports of this commodity from all other countries are subject to unit price lower and upper bounds of $3000 and $8000, respectively. The bounds for Ireland are the country-specific bounds, while the bounds for all other countries are the basket bounds. For the time being, we are only focusing on updating the basket ranges. We will address the issue of multiple ranges later in our discussion of future analyses. We remove from our analysis all commodities that have an interquartile range equal to 0. In order for the resistant fences method to create usable tolerances, the first and third quartiles must be unique values. Otherwise, the lower and upper bounds that will be created will be the same value (a single point, rather than an interval), which is uninformative. Additionally, we removed from our data any commodity that had fewer than 30 records in one year, as more records in a distribution help to prevent outliers from being included in the computation of the quartiles. The sample size requirements are important because the resistant fences procedure is outlier resistant, and the "resistance" of the procedure breaks down when the actual number of outliers exceeds the number of observations in one (or both) of the quartile tails. Larger sample sizes help minimize the probability of breakdown. As a result, we are able to perform our analysis for 11,574 import commodities.

After refining the master file of data, we merged it with the edit master. As mentioned earlier, the edit master, in addition to housing our edit parameters, also contains instructions for how to compute the ratios. For certain commodities, a quantity is not required, and thus, no unit price edit would be performed. For other commodities, quantity may be reported in two different units, and the edit master would provide instructions on which quantity measurement should be used for computing unit price.

Once these files were combined, we wrote a program to create new parameters using the resistant fences method. This program first determines whether a quantity measure is required. If so, it then looks for instructions from the edit master on which measure of quantity to use in computing the unit price. Next, it transforms our data using the natural logarithm transformation. The program then computes the quartiles for each commodity's log-transformed unit price, creates the resistant fences using the appropriate equation for each value of k, and then applies the inverse transformation to the data and newly created fences, bringing everything back to the original scale and allowing us to determine the number of records falling outside the new fences. The end result is a file containing the old edit master parameters, the new resistant fences parameters for each k-value (i.e., 1.5, 2, and 3), and the number and percentage of records (coverage percentage) contained within the old parameters and each new set of parameters. This file will be the tool that analysts can use to determine whether to automatically update the parameters, or look more closely at the data to manually update the parameters if they wish.

5. Results

We found out early in our research that measuring the effectiveness of the resistant fences methodology to create better new bounds is a difficult task. Ideally, we would like to look at two traditional measurements: 1) hit rate (proportion of records flagged out-of-bounds that are actually "bad"), and 2) type 1 error (proportion of "good" records incorrectly flagged as out-of-bounds). However, since some records may contain reporting mistakes that still look reasonable to us, even though they truly are incorrect, or "bad", we have no way of accurately identifying a truly "good" or truly "bad" record. Thus, we need to employ a different type of measurement. We chose, instead, to focus more on issues of coverage, i.e., the percentage of records that fall within the lower and upper bounds. Coverage is an important measure of the analyst workload. With roughly 11,500 commodities for which we can use the resistant fences methodology to create new bounds, it would be very intensive and time-consuming to analyze each individual commodity. Thus, we desire a more

general picture of how the new methodology compares to the existing parameters in our edit

Table 1. Coverage Percentages by Commodity Frequency - Symmetrized Data

New Fences (k=1.5, 2, 3)					
Old (EM)	0 - <50%	50 – <80%	80 – <100%	100%	**Total**
0 - <50%	1	21	149	51	**222**
	1	13	135	73	
	1	7	112	102	
50 – <80%	1	44	473	205	**723**
	1	29	404	289	
	1	17	315	390	
80 – <100%	0	384	7967	733	**9084**
	0	222	7594	1268	
	0	117	6679	228	
100%	0	95	1091	359	**1545**
	0	61	951	533	
	0	37	745	763	
Total	**2**	**544**	**9680**	**1348**	**11574**
	2	**325**	**9084**	**2163**	
	2	**178**	**7851**	**3543**	

master. In Table 1, we categorize each commodity into groups based on their old and new coverage. The rows of this table depict coverage based on the old edit master (EM) bounds, and the columns represent the coverage based on the new bounds. These new bounds were produced using the symmetrized resistant fences methodology with *k*-value of 1.5, 2, and 3, respectively, and only for those commodities with at least 30 records for the entire year. The numbers in the cells depict the number of commodities falling into each category, using the three different *k*-values. Our two primary concerns with edit limits are bounds that are too narrow, such that we have an unusually large number of rejects, and bounds that are too wide, where all records pass through. Bounds that have 0% to less than 50% coverage are far too narrow, and we would like to see the number of commodities with this coverage decrease drastically. We would also like to see a decrease in those commodities whose bounds have 50% to less than 80% coverage, although these bounds are not nearly as narrow as those with lower coverage. Commodities with 100% coverage have bounds that are typically very wide. Thus, we would like to see a reduction in the number of these commodities as well. Commodities with 80% to less than 100% coverage are what we consider to be ideal, and thus, we would like to see an increase in the number of these commodities. The data in Table 1 illustrate that, using our old edit master parameters, we had 222 commodities with coverage less than 50%. The resistant fences method improved this total greatly to only 2 commodities. The number of commodities with 50% to less than 80% coverage also decreases for each *k*-value, and the decrease is greater for larger values of *k*. The tradeoff, however, is that as *k* increases, so do the number of commodities with 100% coverage. Thus, while our old edit parameters yielded 1,545 commodities with 100% coverage, resistant fences brought this total down to 1,348 for *k*=1.5, but actually increased the total for larger *k*-values. We still expect there to be a large number of commodities whose parameters yield 100% coverage, as many commodities rarely have data that is misreported, and thus, no outliers to detect. However, with the resistant fences methodology, we expected this number to be quite a bit smaller than with the edit master fences. This caused us to look at some of these commodities a little closer. We discovered that, in many cases, the parameters produced using the resistant fences method were actually much wider than the edit master limits, but the data did not suggest that wider limits were necessary. We discovered that many commodities' distributions were not statistically skewed, and by log-transforming the data and then performing the inverse transformation, we were actually creating upper bounds that were unreasonably large. Thompson (1999) recommends that symmetrizing data is unnecessary for mildly skewed (skewness < 6.76) distributions, and should never be performed for data with less than 50 observations. As a result, we decided that we would only symmetrize those commodities with at least 50 records and which are highly skewed (skewness >= 6.76). One of the disadvantages of not symmetrizing the data is that many of the new lower bounds would be negative, but in these situations, we simply set the lower bound to be the minimum value in the distribution, so that all lower limits would be greater than 0.

In Table 2, we show how our results change when we only symmetrize the commodities with 50 or more records and high skewness. There are still 2 commodities with less than 50% coverage, but the number of commodities with 100% coverage (i.e., nothing ever rejects), has reduced considerably, to 413 using *k*=1.5. We also noticed that, by using *k*=2, we can cut the number of commodities with 50% to less than 80% coverage and those with 100% coverage approximately in half. At the same time, we increase the number of commodities with the desired 80% to less than 100% coverage to 10,438 (roughly 750 more commodities than the best-case scenario for the all-symmetrized data).

Another result that we observed with both the completely symmetrized data and the "mixed" method was that there were many commodities

Table 2. Coverage Percentages by Commodity Frequency – Mixed (Symmetrized and Unsymmetrized)

New Fences (k=1.5, 2, 3)					
Old (EM)	0 - <50%	50 – <80%	80 – <100%	100%	**Total**
0 - <50%	1	20	177	24	**222**
	1	18	172	31	
	1	11	164	46	
50 – <80%	1	65	603	54	**723**
	1	49	583	90	
	1	39	528	155	
80 – <100%	0	427	8462	195	**9084**
	0	252	8432	400	
	0	131	8002	951	
100%	0	111	1294	140	**1545**
	0	68	1251	226	
	0	48	1132	365	
Total	**2**	**623**	**10536**	**413**	**11574**
	2	**387**	**10438**	**747**	
	2	**229**	**9826**	**1517**	

which, using the old edit master parameters, had the desired coverage of 80% to less than 100%, but whose coverages decreased to less than 80% using the resistant fences methodology. We decided to look closer at some of these commodities, and found that there were many cases where the majority of the data is very similar, and there were few really obvious outliers. In these cases, the quartiles were not equal, but were very similar, such that the new ranges that are produced were too narrow, even for larger values of *k*, and were incorrectly flagging reasonable unit price values as outliers, when they were accepted by the old edit master limits. For these commodities, and for those with 100% coverage and very wide limits, we would want to provide a flag to the analysts. This flag would indicate that changes might be appropriate, and would give the analysts the ability to decide whether to accept the newly suggested limits, or to look at the data and decide if other limits would be more appropriate.

6. Conclusions

When we initially symmetrized all the data, we were somewhat concerned about the number of very wide limits we were observing. These wide limits correspond to those commodities whose new limits yield 100% coverage. We do not wish to eliminate all commodities with 100% coverage, but when we looked more closely at some of these commodities, we noticed that there were many situations where the new upper limits were extremely large. This was due to the inverse natural log transformation combined with sample size and skewness coefficient limitations. Once we decided not to symmetrize all the data, we were able to decrease the number of very wide limits considerably.

Ideally, we would like a new methodology to create new limits whose coverages are primarily in the desired 80% to less than 100% range. Although we are not concerned with some commodities having 100% coverage, we would like the number of these commodities to be less than what we have now with the current edit master parameters, and we would also like the number of commodities with low (<50%) coverage to decrease considerably. In order to achieve this balance, it appears that the methodology that works best on import unit price edits is the mixed method, where we symmetrize the data with at least 50 records and with skewness at least 6.76, and do not symmetrize the others, with a *k*-value of 2. With this method, we get the most commodities with the desired coverage, a drastic reduction in those with 100% coverage, and not too many with less than 80% coverage. Yet, while this methodology seems to work the best of those we tested, we still intend to examine other ways to make it even more powerful, and will discuss some of these in the next section.

7. Further Analysis

In order to get an automated system of updating edit master parameters that our analysts are comfortable with, there are several other issues that need further investigation.

First of all, the results we have produced so far show how coverages change when we use the new methodology as opposed to what our existing parameters provide. These results do not necessarily reflect the level of coverage that we would experience upon implementing a new methodology, because it is unlikely that we would change the parameters for all commodities each time an update is run. More than likely, we will need to determine threshold levels to distinguish 'substantially large' changes to our parameters from 'reasonable' changes. Those new parameter values that may reflect 'substantially large' changes will, more likely, be used to flag situations where an analyst needs to further examine the data, and determine whether a change is appropriate. Developing threshold levels requires some investigation, and assistance from subject-matter analysts. Once we have some decision logic in place, and can fully develop a set of updated parameters, we can then test future months of data on these edit parameters to determine their effectiveness.

In our analysis thus far, we looked at some results when we symmetrize all the data using the

natural log transformation, and when we combine symmetrized data with unsymmetrized commodities. We then looked at these results for different *k*-values, but did not examine further if results could be improved by using different *k*-values for the symmetrized and unsymmetrized data. For example, we may be able to further increase our coverage figures if we use a *k*-value of 1.5 for the symmetrized data, and a *k*-value of 2 for the unsymmetrized data. It may prove valuable to try different *k*-value combinations to maximize our results.

Also, we have based our research solely on import unit price edits. We need to further test the resistant fences methodology, not only on other import ratio edits, but on all export ratio edits as well.

Another very important aspect of our future research involves the creation of country-specific parameter values. As we mentioned earlier, our current analysis only handles 'basket ranges' and does not attempt to create country-specific ranges. Research may show that we can do a better job of editing our data if we can update the country-specific ranges, and also, if we allow the countries with specific ranges to change, based on changes in trade patterns. This would further improve our overall coverage.

Country-specific ranges are not the only "sub-ranges" that may be possible. As improvements in technology and computing ability become greater, we are also exploring new ways of analyzing the data. One idea that is being studied is the editing of data based on companies. For large companies with enough data on a regular basis, it may be possible to create sets of reasonable parameters to specifically edit the data submitted by those companies. An automated system of using historical data to create parameters, like the resistant fences methodology, will be crucial to development and testing of these, or any similar, "new" edit parameters.

Finally, we would also like to be able to use this methodology for as many commodities as possible. Recall that commodities are classified using 10-digit commodity codes, and only those commodities with at least 30 records are eligible. It might be possible, through additional research, to combine together some of the commodities that currently do not have enough data. Perhaps, by combining such commodities at the 8-digit level, or even the 6-digit level, we can obtain enough data to create new limits for the grouping of commodities, and thus, increase the number of commodities whose parameters can be updated using this resistant fences methodology.

8. References

Thompson, Katherine J. and Sigman, Richard S. (1999). Statistical Methods for Developing Ratio Edit Tolerances for Economic Data. *Journal of Official Statistics*. Vol. 15, No.4, pp. 517-535

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

Tukey, John W. (1976). Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley.

ACKNOWLEDGEMENTS

The authors would like to thank Katherine Jenny Thompson, Maria Garcia, Paul Herrick and Renee M. Coley for providing assistance in this work and helpful comments on the paper.