

Risk analysis approaches to rank outliers in trade data

Vytis Kopustinskas, Spyros Arsenis

European Commission, Joint Research Center

Institute for Protection and Security of Citizen, Via E.Fermi 2748, Ispra (VA), Italy

Email: vytis.kopustinskas@jrc.it; spyros.arsenis@jrc.it

Abstract: The paper discusses the trade data outliers ranking methods based on statistical information with the objective to prioritize anti-fraud investigation activities. The paper presents a ranking method based on risk analysis framework and discusses more complex trade fraud indicator that aggregates a number of individual statistical criteria.

Keywords: risk analysis; outliers detection; trade data, ranking and prioritisation

1. Introduction

The detection of outliers in trade data can be important for various practical applications, in particular for customs fraud prevention or data quality analysis. From customs inspector point of view, the trade transactions classified as outliers might be of interests due to possible on-going fraud activities. For example, low price outliers might indicate that the specific transaction is undervalued to evade import duties. Another example is trade based money laundering that might be indicated by both low and high price outliers.

The statistical algorithms used to detect outliers in large trade datasets typically produce high number of transactions classified as outliers [1,2]. From customs inspector perspective, high number of transactions flagged as suspicious is difficult to handle. Therefore the detected outliers must be ranked according to certain criteria in order to prioritize the investigation actions. Different criteria could be used for ranking purposes and they are derived from at least two very distinct information sources: statistical information about the data and customs in-house information system.

The paper discusses outliers ranking methods based on statistical information with the objective to prioritize anti-fraud investigation actions. Although the ranking based on solely statistical information cannot be considered complete, as in-house customs information is also important, statistical properties of the dataset represent a valuable source of information that can be used to rank outliers.

2. Risk analysis framework

Risk analysis framework can be applied to the trade data outliers ranking problem. The fundamental questions in quantitative risk analysis are the following [3]:

1. What can go wrong?
2. How likely it will happen?
3. If it happens, what consequences are expected?

To answer question 1, a list of initiating events should be defined. The likelihood of the events should be estimated and the consequences of each scenario should be assessed. Therefore, quantitatively risk can be defined as the following set of triplets:

$$R = \langle S_i, P_i, C_i \rangle, i = 1, \dots, n. \quad (1)$$

where S_i - scenario of the initiating events; P_i - likelihood of the scenario; C_i - consequence of the scenario; n - number of scenarios.

In case of outliers in trade data, the triplet can be interpreted in the following way: P – likelihood that an outlier is a real fraudulent transaction; C – consequence of the fraudulent trade transaction (e.g. unpaid taxes or duties). The interpretation of S can be important only if several methods are used to detect outliers.

3. Approaches to rank low price outliers in trade data

There are several approaches to obtain numerical ranking based on risk analysis framework. In trade data application the most suitable method is to multiply P and C , where C is an estimate of the loss to the budget (unpaid duties) and P is likelihood that specific transaction is a fraud. The multiplication $R = P \times C$ provides an average fraud related damage estimate caused by specific trade activity and allows comparing different outliers with the purpose to rank them according to their severity.

The Figure 1 gives an illustration of low price outliers detected in a trade dataset. In order to use the risk analysis framework (1) we have to estimate the likelihood (P) - probability of fraud in the trade transaction. This is not an easy quantity to estimate, but using statistical information we can assume that p-value produced by statistical tests for outliers could be a suitable approximation. It means that the lower the p-value, the more likely it is a fraudulent transaction. In practice, this can also be a mistype error and not a fraud, but from customs officer point of view, the low price outliers are suspicious and should be investigated.

As the relationship between P and p-value is reverse, the transformation is used:

$$\begin{cases} P = \frac{-\log_{10}(pvalue)}{10}, & \text{if } pvalue \geq 10^{-10} \\ P = 1, & \text{if } pvalue < 10^{-10}. \end{cases} \quad (2)$$

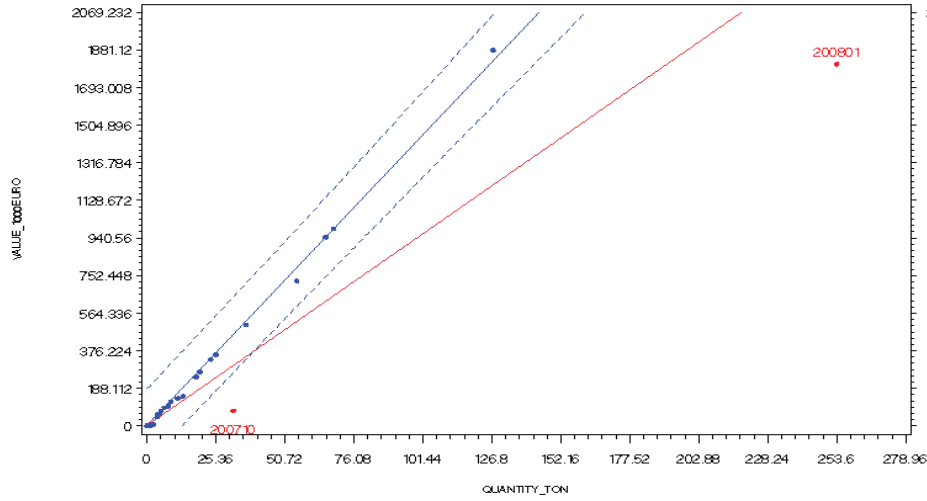


Figure 1: The detected outliers (red color) in an example trade dataset for specific Product/Origin/Destination. The angle of red and blue lines indicates the average unit price from all data (red) and after removal of the outliers (blue).

By transformation (2) the p-value is transformed into scale $[0,1]$, and the effect is limited to 1. The scale here is arbitrary and chosen mainly for the purpose of convenience, driven by the fact that extremely low p-values are no more informative for the ranking purposes.

The consequence part (C) of (1) can be estimated by multiplying traded quantity (Q) and transaction unit price difference (ΔU) from the recorded to the estimated ‘fair’ price: $C = Q \times \Delta U = Q \times (\hat{U} - U)$, where $\hat{U}$ – the estimated ‘fair’ transaction unit price determined by the regression with removed outliers; U – the unit price as recorded ($U = V/Q$); V – value of the transaction as recorded. The interpretation of C is an average loss to the budget if the underlying transaction is in fact fraudulent.

Thus the trade fraud indicator (R) can be computed as follows: $R = P \times Q \times \Delta U$. The indicator can also be transformed into $[0,10]$ scale by using min-max transformation, as this scale is more self-evident to investigators. The investigator should start the investigation process from the outlying trade transactions with the highest values of R .

The R is a simple and easy to understand indicator, however the trade dataset contains much more statistical information. Table 1 provides a number of criteria which could be used for the development of more complex ranking structure.

Table 1: Statistical criteria for the development of ranking structure, indicating their original scale and rescaling method. V_{PO} – Trade value by aggregating all destinations Q_{PO} – Trade quantity by aggregating all destinations.

No	Criteria	Original scale	Rescaling
I_1	Quantity traded Q	$[0,\infty]$	Min-max transformation to $[0,1]$
I_2	Value traded, V	$[0,\infty]$	Min-max transformation to $[0,1]$

I_3	Average loss $Q \times \Delta U$	$[0, \infty]$	Min-max transformation to $[0, 1]$
I_4	Ratio $\hat{U}/U$	$[0, \infty]$	Min-max transformation to $[0, 1]$
I_5	Ratio V/V_{PO}	$[0, 1]$	No
I_6	Ratio Q/Q_{PO}	$[0, 1]$	No
I_7	Number of obs/36	$[0, 1]$	No
I_8	P -value	$[0, 0.1]$	Log transformation to $[0, 1]$ as in (2)
I_9	Coeff. of determination R^2	$[0, 1]$	No
I_{10}	Final R^2 /Initial R^2	$[0, \infty]$	Min-max transformation to $[0, 1]$

The criteria listed in Table 1 are all numerical and their higher value is associated with the higher impact to trade fraud indicator (FI). After rescaling as shown in Table 1, the individual criteria (I_i) are comparable among themselves. The specific weights (w_i) must be assigned to each criterion to determine its relative impact to the final indicator score. The most popular method to combine different criteria into a single numerical indicator is to compute a weighted sum: $FI = \sum_{i=1}^m w_i \times I_i$, where m – number of individual criteria.

The complication arises from the fact that some criteria could be highly correlated and therefore correlation matrix must be thoroughly examined before assignment of weights to each criterion. Without prior knowledge, equal weights could be assigned to non-correlated criteria. Having opinion of subject matter experts, the weights can be estimated by analytic hierarchy process (AHP) which provides a comprehensive and rational framework for structuring decision problems.

4. Final remarks

The paper discusses the trade data outliers ranking methods with the objective to prioritize anti-fraud investigation actions. The risk analysis framework was used to develop a ranking structure. The more complex trade fraud indicator is also discussed that consists of a number of individual statistical criteria. The final version of the paper will include demonstration and comparison of different trade fraud indicators.

References

1. D. Perrotta and F. Torti. Detecting price outliers in European trade data with the forward search. In *Springer Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, 2009. To appear.
2. Atkinson A.C., Riani M., Cerioli A. (2004) *Exploring Multivariate Data With the Forward Search*, Springer Verlag, New York.
3. Kaplan, S., Garrick, B. J. (1981) On the quantitative definition of risk, *Risk Analysis*, 1(1), 11-27.