UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Editing

7-9 October 2024, Vienna

Detecting Extreme Numerical Outliers in Trade Data: A Novel Method for Highly Asymmetric Distributions

Andrea Cerasa (European Commission, Joint Research Centre)

andrea.cerasa@ec.europa.eu

I. INTRODUCTION

1. The accurate detection of extreme numerical outliers in trade data is crucial for effective policy making, anti-fraud measures, and reliable EU-wide statistics. The European Commission could significantly benefit from the Surveillance database of DG TAXUD, which collects import/export transactions from national authorities. However, large errors in declared values may occur due to data quality issues, and this can severely impact data analyses and lead to incorrect decisions[Perrotta et al., 2020].

2. The significant skewness observed in the distributions of international trade data poses a major challenge in identifying these extreme numerical outliers, as it can be difficult to distinguish them from the normal values in the right tail of the distribution. Building upon the work described in the scientific article "An adjusted boxplot for skewed distributions" [Hubert and Vandervieren, 2008], our proposed method addresses the challenges posed by the highly skewed distributions found in international trade data. In particular, the primary contributions of our proposal are:

(i) Enlarging the spectrum of distributions considered in the article to more closely resemble the asymmetric distributions that characterize international trade data. This allows for a more accurate representation of trade data distributions, which are often more skewed than those found in the existing literature.

(ii) Developing a method for calculating thresholds that identify extreme anomalous numbers of each distribution, rather than adapting a box plot for skewed distributions. This novel approach is specifically designed for detecting extreme numerical outliers in highly asymmetric distributions.

3. To assess its quality, we will test it on real international trade data provided by DG TAXUD. This practical deployment will enable us to evaluate the effectiveness of our approach in detecting extreme numerical outliers in trade data and contribute to the improvement of data quality checks at DG TAXUD. It will also assess the potential to significantly enhance the reliability of EU-wide statistics, anti-fraud measures, and policy making, as well as facilitate economic operators in their activities.

4. This paper is organized as follows. Section II describes the motivation and the statistical challenges related to the problem under study. Section III presents the statistical approach. Section IV discusses the empirical results obtained with Surveillance data. Finally, Section V concludes.

II. MOTIVATION AND STATISTICAL CHALLENGES

1. The European Union (EU) has created a robust legal structure designed to facilitate efficient customs operations among its member states through the Union Customs Code (UCC) - Regulation (EU) No 952/2013, along with its implementing (EU 2015/2447) and delegated (EU 2015/2446) regulations. This framework is crucial for streamlining customs procedures, preventing fraud, securing the EU external borders, and facilitating legitimate trade. The information, collected on a daily basis from the national authorities, form the Surveillance database of the Directorate-General for Taxation and Customs Union (DG TAXUD). Each entry of this database corresponds to an import or export transaction, and contains information recorded by the trade operators in a customs declaration, including the net and gross mass of the product traded, the corresponding economic value, the origin and the destination of the consignment. The TARif Intégré Communautaire (TARIC, Integrated Tariff of the European Communities) database, which represents the EU's comprehensive customs tariff, categorizes products using a hierarchical coding system. The Surveillance and TARIC systems facilitate the monitoring of EU trade for different purposes: policy-making, compiling EU-wide statistics, ensuring the integrity of supply chains, combating frauds, and supporting economic operators with tasks such as identifying licensing needs or calculating duties. In 2022, for example, these systems were instrumental in monitoring trade related to the EU's response to the Russian invasion of Ukraine, aiming to swiftly identify transactions involving sanctioned goods destined for Russia or Belarus to enable robust export controls by EU Customs.

2. As expected considering the quantity of single EU imports and exports that occurs everyday, the amount of import and export records stored in Surveillance is considerably huge. In 2023 alone, 990,937,480 entries were recorded. This massive volume of data, combined with the fact that their transmission typically occurs shortly after the actual completion of import/export transactions, makes the presence in Surveillance data of excessively outlying values a significant concern. Such outliers, typically unintended,

can greatly distort the figures declared at customs — most notably affecting the mass, statistical value, or number of supplementary units — and can consequently impact data analyses, lead to incorrect conclusions, and result in poor decision-making. It is therefore imperative for DG TAXUD to implement dependable data quality checks before the Surveillance data is used.

3. At first sight, one might consider classical robust statistical methods for detecting extreme outliers in data[1]. For example, without making distributional assumptions that could be violated by Surveillance data, we might analyze all records for a specific product, focusing on a particular numerical variable (e.g. net mass), and establish a threshold based on a significant deviation from the median, expressed in terms of Median Absolute Deviations (MAD) or the Interquartile Range (IQR)[2]. Alternatively, one might use the standard boxplot method (see Figure 1). However, as Figure 2 demonstrates, these methods may not yield reliable results when applied to highly skewed distributions, which are common for many products in Surveillance database. Even though the practical case in the figure does not contain any extreme value, both approaches flag numerous potential outliers, even with the most conservative threshold $\text{Median} + 5 \times 1.4826 \times \text{MAD}$.

4. It is evident that using the two classical robust statistical methods on Surveillance data implies a significant risk of over-declaring the relevant anomalies. This motivates our study for an alternative approach that, besides guaranteeing statistically solid outcomes, should ideally take into account the following requirements:

a) Flexibility: the proposed method should be suitable for all the products in the database.
b) Statistical properties: the statistical approach should be easy to apply and to explain, allowing for a strict control of the false alarms. This can go at the expense of the ability to detect all anomalies, but we aim at providing to the Customs offices only a manageable set of irrefutable errors.
c) Computational efficiency: the procedure detects new outliers every day on datasets that could be quite large, and needs to be fast.
d) Software simplicity: the approach should be based on algorithms that run in an Oracle database with limited calls to specialised statistical functions.

III. DESCRIPTION OF THE STATISTICAL APPROACH

1. The boxplot, originally proposed by Tukey et al. [1977], is one of the most frequently used graphical techniques for visualizing the distribution of continuous unimodal data. For a univariate dataset $X \equiv \{x_1, x_2, \dots, x_n\}$, it combines information about the

[1] Robust statistical methods aim to provide reliable and accurate results even when the data contain outliers.

[2] For a univariate dataset $X \equiv \{x_1, x_2, \dots, x_n\}$, the MAD is defined as $\text{Median}(|x_i - \text{Median}(X)|)$, whereas the IQR as the difference between the third quartile Q_3 (i.e. the 75th percentile) and the first quartile Q_1 (i.e. the 25th percentile).

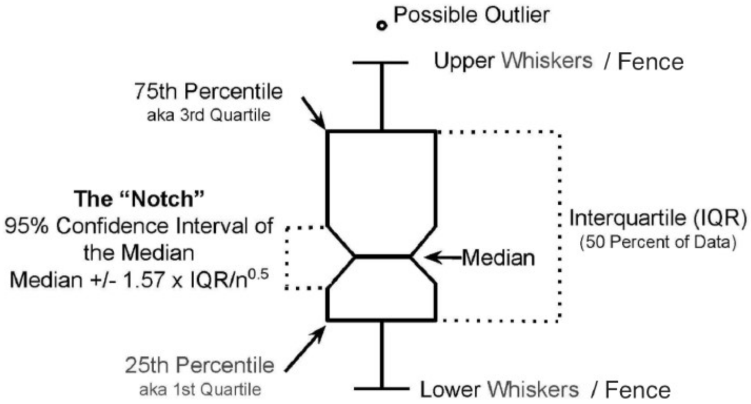

FIGURE 1. Graphical representation of a boxplot.

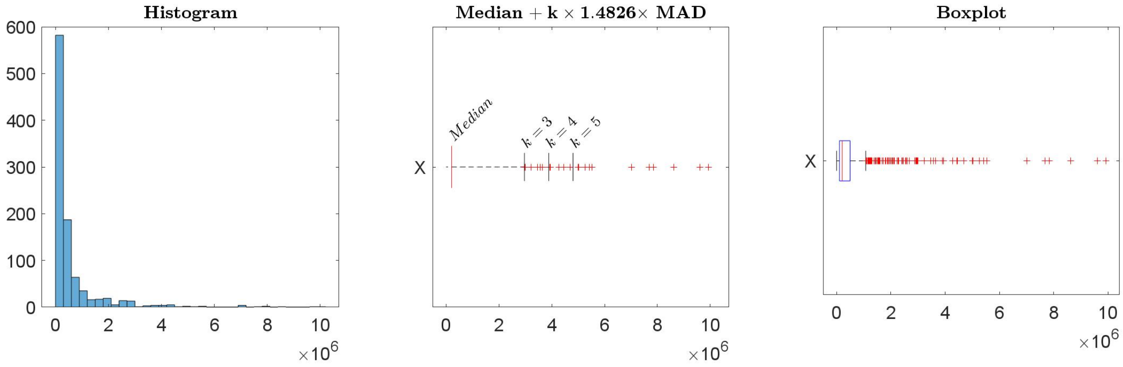

FIGURE 2. Application of classical robust methods on a Surveillance dataset

location, spread, skewness and tails of the data in a very intuitive and easily understandable manner (see figure 1). It also allows for flagging as *potential outliers* those points which are outside the fence boundaries defined as:

$$[Q_1 - 1.5 \text{ IQR}; \; Q_3 + 1.5 \text{ IQR}]$$

Observations outside these thresholds are not necessarily outliers [Hoaglin et al., 2000]. If the data are normally distributed, approximately 0.7% of the observations are expected to lie outside the fence (about 0.35% in each tail of the distribution). For distributions with thicker or thinner tails than the normal distribution, this percentage is expected to be larger or smaller, respectively. Similarly, for skewed distributions, we expect a different number of potential outliers on each side of the distribution, depending on the direction and degree of the skewness. For example, applying the boxplot method to the dataset represented in Figure 1 results in 0% of potential outliers in the left tail and 12.59% in the right tail.

2. Therefore, while the traditional boxplot is an invaluable tool for data analysis, it has limitations when applied to distributions that are highly skewed, such as those commonly encountered in international trade data. The symmetric nature of the fences does not account for the intrinsic asymmetry of the data, leading to a significant risk of over-declaring the relevant anomalies. Recognizing this issue, Hubert and Vandervieren [2008] proposed an adjustment of the traditional boxplot where the fences are based on a measure of skewness, allowing for asymmetric whisker lengths that better reflect the underlying distribution of the data. In particular, they defined the fences as:

$$[Q_1 - h_L(\mathrm{MC})\ \mathrm{IQR};\ Q_3 + h_U(\mathrm{MC})\ \mathrm{IQR}]$$

with $h_L(\cdot)$ and $= h_U(\cdot)$ are two different functions of the skewness, measured by the medcouple introduced in Brys et al. [2004]:

$$\mathrm{MC} = \underset{x_i \le Q_2 \le x_j}{\mathrm{median}} \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}.$$

It follows from this definition that $-1 \le \mathrm{MC} \le 1$, and that $\mathrm{MC} = 0$ when the distribution is symmetric. Hubert and Vandervieren [2008] considered several specifications for $h_L(\cdot)$ and $h_U(\cdot)$, all with the constraint $h_L(0) = h_U(0) = 1.5$ in order to obtain the standard fence expression in case of symmetric distributions. The specification with the most desirable properties resulted to be the exponential one:

$$h_L(\mathrm{MC}) = 1.5\mathrm{e}^{a\,\mathrm{MC}} \qquad h_U(\mathrm{MC}) = 1.5\mathrm{e}^{b\,\mathrm{MC}}$$

with $a = -3.79$ and $b = 3.87$. These two values were obtained by estimating the following linear regression without intercept:

$$\begin{cases} \ln\left(\dfrac{2}{3}\dfrac{Q_1 - Q_{0.35\%}}{\mathrm{IQR}}\right) \approx a\ \mathrm{MC} \\ \ln\left(\dfrac{2}{3}\dfrac{Q_{99.65\%} - Q_3}{\mathrm{IQR}}\right) \approx b\ \mathrm{MC} \end{cases} \tag{1}$$

using the values of $Q_1, Q_3, Q_{0.35\%}, Q_{99.65\%}$ and MC calculated on samples of 10,000 observations extracted from 12,605 distributions coming from the family of Γ, χ^2, F, Pareto and G_g. The parameter space for each family was chosen such that $0 \le \mathrm{MC} \le 0.6$. Finally, the choice 0.35% and 99.65% quantiles aims at replicating the expected 0.7% of marked outliers in the standard boxplot at the normal distribution.

3. The main limitation in applying this approach to datasets extracted from Surveillance is that it was built based on distributions with a moderate degree of asymmetry ($\mathrm{MC} \le 0.6$). However, empirical datasets are characterized by much higher MC values. Figure 3 shows the values of the medcouples calculated for the 7,447 products codes that will be object of our empirical exercise described in Section IV. For almost 80% of them, the value of the medcouple is larger than 0.6. Extending the model to cases of such pronounced asymmetry is not straightforward, as confirmed by Hubert and Vandervieren [2008]: *"It appeared that constructing one good and easy model that also includes the cases with MC > 0.6 is hard, hence we only concentrated on the more common distributions with moderate skewness"*. On the other hand, the goal of our study is less ambitious.

Indeed, our aim is not to extend the classic boxplot to the case of highly asymmetric distributions, but simply to identify for each empirical dataset a threshold value capable of isolating the extreme outliers of the distribution. Furthermore, given the context of our application, we might limit ourselves to consider only the right tail of the distribution, disregarding what happens in the left tail.

4. Considering these factors, we opted for an approach that builds upon the methodology proposed by Hubert and Vandervieren [2008], tailored to meet the specific requirements of our study. First of all, since our primary interest lies in detecting potential extreme outliers within the right tail of the distribution, we can focus exclusively on the corresponding regression, and disregard the regression associated with the left tail. Secondly, since our objective is not to replicate the main features of the standard boxplot, we adopt a more flexible specification of the regression model, that is:

$$\ln(x^i_{max} - Q^i_3) \approx \alpha + \beta \cdot \ln(Q^i_1) + \gamma \cdot \ln(Q^i_2) + \theta \cdot \ln(Q^i_3) + \phi \cdot \ln(Q^i_{95\%}) + \lambda \cdot \mathrm{MC}^i \quad (2)$$

where $i = 1, \ldots, 7,447$. Therefore, the dependent variable and the regressors are not derived from simulated theoretical distributions, but from the empirical observations of Surveillance products. Consequently, the coefficients $\alpha, \beta, \gamma, \theta, \phi, \lambda$ are estimated using the maximum values, percentiles, and medcouples computed from Surveillance datasets. This is why a robust regression technique is preferable for the estimation of (2). The values of the dependent variables may indeed be influenced by the potential presence of the extreme outliers, introducing bias in the estimates. A robust regression method might mitigate this issue [Huber and Ronchetti, 2011, Rousseeuw and Leroy, 2005]. It is important to underline that model (2) embeds the implicit assumption that the values of $Q_{95\%}$ do not contain outliers. This corresponds to assume that the percentage of outliers in each Surveillance product cannot be larger than 5%.

IV. EMPIRICAL RESULTS

1. For the empirical exercise, we focus only on the values of the net mass. Subsequent phases of the study could include additional quantitative variables from the Surveillance database. We considered all Surveillance entries recorded from 01/11/2022 to 29/05/2024. The total number of records for import and export declarations during this period is 1,556,162,485 across 9,816 different products. Not all products were included in the analysis; those with fewer than 100 distinct net mass values were excluded, as were those with a negative medcouple value, since extreme outliers are not expected in products with negative skewness. These filters significantly reduced the number of analyzed products to 7,447 (a reduction of approximately 24%). However, the total number of entries in the filtered products was 1,528,757,402, which represents over 98% of the total. All empirical results presented in this section were obtained using MATLAB routines.

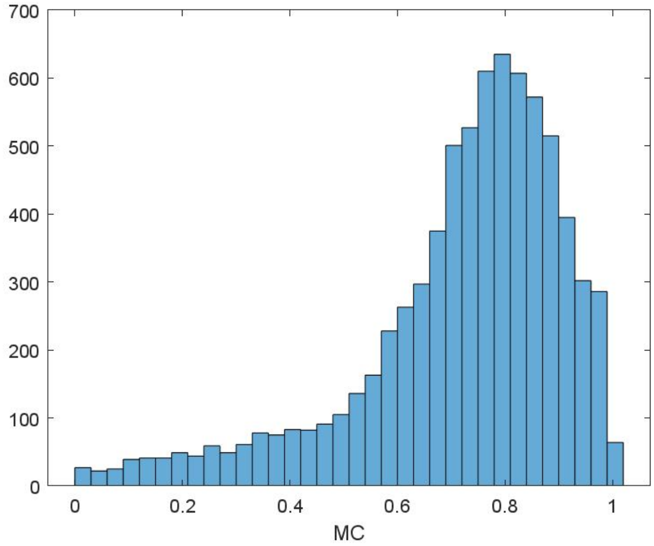

FIGURE 3. Histogram of the medcouple (MC) values calculated on selected products in Surveillance database. The characteristics of the selection are described in Section IV.

2. The objectives of the empirical exercise were as follows: (i) Robust estimation of the coefficients for regression (2) using the statistics calculates on Surveillance datasets; (ii) Use the regression estimates to establish a reliable threshold for each product to flag potential extreme outliers; (iii) Evaluation of the quality of detected extreme outliers through the analysis of practical examples.

A. **Robust estimates**

1. The standard estimation of (2) through Ordinary Least Squares (OLS) might yield biased results due to the potential presence of outliers in the 7,447 values of the dependent variable $\ln(x_{max} - Q_3)$. To mitigate this issue, we employed the following robust regression methods:

- the Iteratively Reweighted Least Squares (IRWLS) [Holland and Welsch, 1977] with the leverage adjustment of residuals suggested by Dumouchel et al. [1989], calculated with the MATLAB function `robustfit`.
- the MM estimator [Maronna et al., 2019] calculated with the MATLAB function `MMreg` included in the FSDA toolbox [Riani et al., 2012].
- the Forward Search (FS) estimator [Riani et al., 2015] calculated with the MATLAB function `FSR` also included in the FSDA toolbox.

Table 1 presents the estimated coefficients, with OLS estimates included for comparison. For both IRWLS and MM, among the multiple options available, the default *bisquare* weight and ρ functions were used, given that variations did not significantly impact the estimates.

2. The consistency in the signs of the estimated coefficients across different estimation methods is notable, with all approaches yielding a positive intercept and positive effects on the dependent variable for all regressors except $\ln(Q_3)$. When comparing OLS estimates with robust methods, the latter suggest stronger effects for $\ln(Q_2)$, $\ln(Q_3)$, $\ln(Q_{95\%})$, and a weaker effect for the constant term, $\ln(Q_1)$, and MC. As anticipated, the robust estimates of the regression standard error σ are remarkably lower than those obtained with OLS. The IRWLS and MM estimates are particularly similar, likely due to their methodological similarities and use of the same kernel function with settings that ensure an asymptotic efficiency of 95% for both. Finally, Figure 4 exhibits the distribution of standardized residuals for all four estimation methods. The plots confirm the presence of a subset of right-tail observations that likely represent the extreme anomalies our study aims to identify.

TABLE 1. Estimates of regression (2) using standard OLS and different robust approaches

	OLS	**IRWLS**	**MM**	**FS**
$\hat{\alpha}$	4.291	4.191	4.192	4.007
$\hat{\beta}$	0.138	0.102	0.102	0.104
$\hat{\gamma}$	0.064	0.088	0.087	0.146
$\hat{\theta}$	-0.332	-0.392	-0.392	-0.468
$\hat{\phi}$	0.959	1.011	1.011	1.028
$\hat{\lambda}$	0.771	0.519	0.517	0.767
$\hat{\sigma}$	1.610	1.167	1.083	0.881

B. The identification of a threshold for flagging extreme outliers

1. With the estimates obtained from model (2) and its standard error, a natural definition of the thresholds for the maximum values of each product can be:

$$\ln(x^i_{max} - Q^i_3) < \hat{\alpha} + \hat{\beta}\cdot\ln(Q^i_1) + \hat{\gamma}\cdot\ln(Q^i_2) + \hat{\theta}\cdot\ln(Q^i_3) + \hat{\phi}\cdot\ln(Q^i_{95\%}) + \hat{\lambda}\cdot\mathrm{MC}^i + k\hat{\sigma}$$

$$\Downarrow$$

$$x^i_{max} < Q^i_3 + e^{\hat{\alpha}+\hat{\beta}\cdot\ln(Q^i_1)+\hat{\gamma}\cdot\ln(Q^i_2)+\hat{\theta}\cdot\ln(Q^i_3)+\hat{\phi}\cdot\ln(Q^i_{95\%})+\hat{\lambda}\cdot\mathrm{MC}^i+k\hat{\sigma}}$$

Observations above this threshold for product i may be flagged as extreme outliers. The multiplier k determines the conservatism of the threshold: higher (smaller) values result in fewer (more) flagged outliers. Figure 5 shows the patterns of the total number of extreme values identified on the 7,447 products through the four models for increasing values of k. The starting value is $k = t_{97.5\%,7441} = 1.96$ (which corresponds to the multiplier one would use for building a 5% confidence interval for the dependent variable),

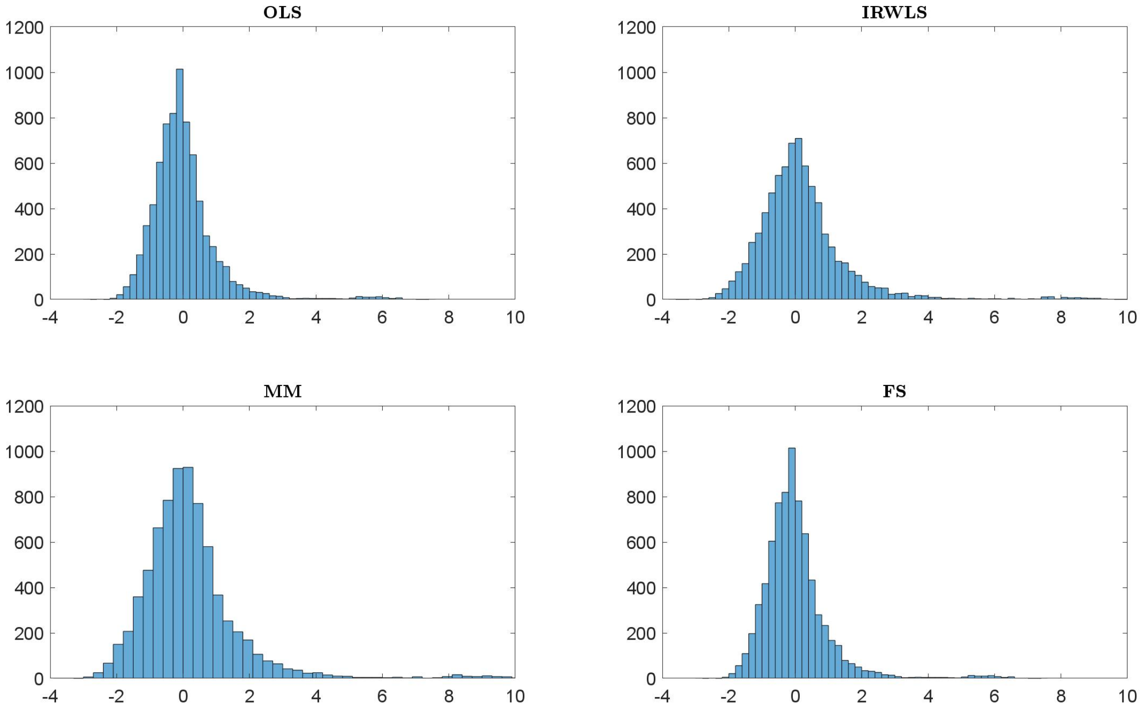

Figure 4. Standardized regression residuals

whereas the maximum is $k = 5$. As expected, the number of flagged observations decreases with increasing k. Comparing the patterns of the four methods, it is clear that for a fixed value of k the larger $\hat{\sigma}$, the smaller the number of spotted observations. The two extremes are the FS, that starts with 5,475 declared outliers (accounting for 0.00036% of the total number of records) and ends with 336 (0.00002%), and the OLS, that starts with 667 (0.00004%) and ends with 143 (0.00001%). The intermediate outcomes offered by the other two methods seem to be a balanced compromise.

2. Without a deterministic rule for the optimal k value, an alternative approach is to calculate for each observation x_j^i exceeding Q_3^i the following value:

$$\kappa_j^i = \frac{ln(x_j^i - Q_3^i) - \hat{\alpha} - \hat{\beta} \cdot \ln(Q_1^i) - \hat{\gamma} \cdot \ln(Q_2^i) - \hat{\theta} \cdot \ln(Q_3^i) - \hat{\phi} \cdot \ln(Q_{95\%}^i) - \hat{\lambda} \cdot \mathrm{MC}^i}{\hat{\sigma}}.$$

In practice, κ_j^i represents the minimum k required to avoid flagging x_j^i as an outlier. In other words, if $\kappa_j^i = 4$, then x_j^i is not an outlier for $k \leq 4$, whereas it is flagged when $k > 4$.

Figure 6 illustrates the pattern of ordered κ_j^i values, limited to cases where $\kappa_j^i > 1.96$. The four lines share similar characteristics, with a point of inflection where the slope of the line increases, suggesting a potential empirical threshold. This point corresponds to κ_j^i values of approximately 2.75, 4, 4.25, and 5.5 for OLS, IRWLS, MM, and FS, respectively. With these thresholds, the number of extreme values flagged by the four methods is approximately 250. Small differences were registered only for 15 products. This means that 99.8% of the times, the four methods provide the same outcome. Moreover, more than 98% of time the common result is "no extreme observations to flag".

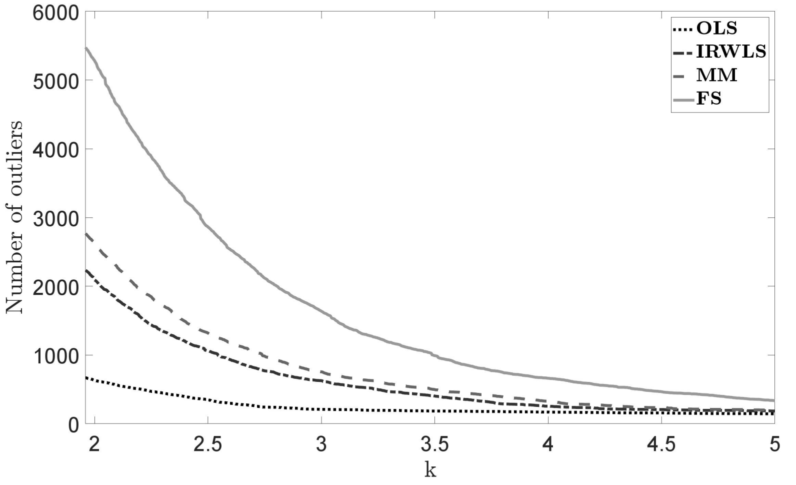

FIGURE 5. Number of flagged outliers for increasing values of k.

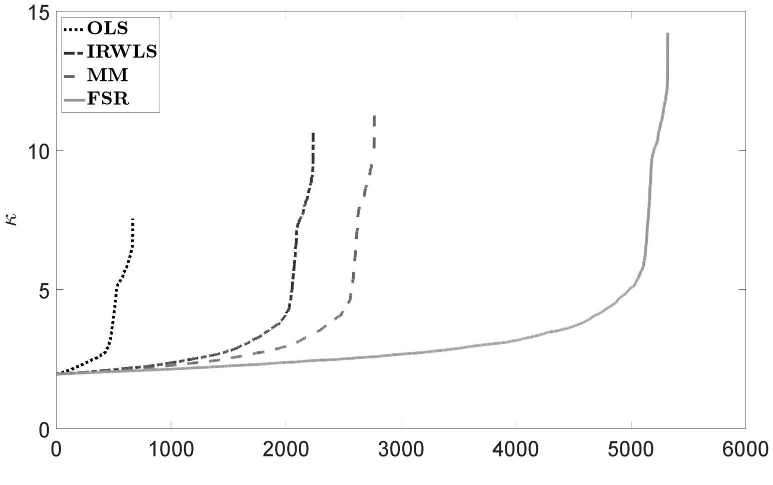

FIGURE 6. Patterns of the ordered values of κ_j^i.

C. Some practical examples

Figure 7 presents the results of the proposed outlier identification strategy applied to three Surveillance products. In all cases, the four methods agreed on the outcomes. The comparison of the scales of the upper and lower panels highlights the magnitude differences between flagged and non-flagged observations. This discrepancy significantly affects the representation of non-flagged observations, which appear compressed and overlapping near zero in the upper panels, while in the lower panels, they are well-spaced and distinguishable.

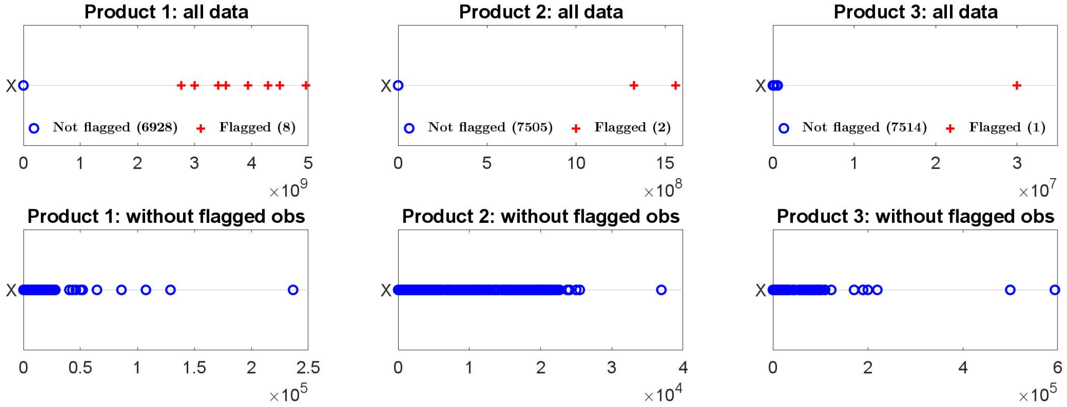

FIGURE 7. Examples of extreme values detection on Surveillance product.

V. CONCLUSIONS

1. This study set out to address the challenge of detecting extreme numerical outliers in the highly skewed distributions typically found in international trade data. By expanding upon the adjusted boxplot methodology for mildly skewed distributions developed by Hubert and Vandervieren [2008], we have proposed a robust statistical approach suitable for the highly skewed distribution observed in real-world trade data. Through robust regression analysis, we have successfully estimated a model that provides tailored thresholds for flagging extreme numerical outliers in each product. The empirical application of our approach has demonstrated its effectiveness in accurately identifying extreme outliers. Our method also meets the specific needs in terms of flexibility, statistical robustness, computational efficiency, and software simplicity.

2. By ensuring a more reliable outlier detection mechanism, this study contributes significantly to the integrity of international trade data analysis. The implications are wide-ranging, as better data quality checks can potentially enhance the reliability of

EU-wide statistics, the effectiveness of anti-fraud measures, and the accuracy of policy-making decisions. Moreover, economic operators may benefit from improved data quality by obtaining clearer insights into their trade activities.

3. In conclusion, the methods developed in this study serve as a robust and adaptable tool for addressing one of the key challenges in statistical analysis of international trade data. Their integration with current approaches is ongoing [Perrotta et al., 2023, available upon request]. Future work could extend this approach to additional quantitative variables within the Surveillance database, in order to further refine the process of outlier detection and increase the robustness of trade data analyses.

References

Guy Brys, Mia Hubert, and Anja Struyf. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13(4):996–1017, 2004.

William Dumouchel, Fanny O'brien, et al. Integrating a robust option into a multiple regression computing environment. In *Computer science and statistics: Proceedings of the 21st symposium on the interface*, pages 297–302. American Statistical Association Alexandria, VA, 1989.

David C Hoaglin, Frederick Mosteller, and John W Tukey. *Understanding robust and exploratory data analysis*, volume 76. John Wiley & Sons, 2000.

Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

Mia Hubert and Ellen Vandervieren. An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, 52(12):5186–5201, 2008.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

Domenico Perrotta, Enrico Checchi, Francesca Torti, Andrea Cerasa, and Xavier Arnes Novau. Addressing price and weight heterogeneity and extreme outliers in surveillance data - the case of face masks. *Publications Office of the European Union*, JRC122315, 2020. doi: DOI:10.2760/817681.

Domenico Perrotta, Enrico Checchi, Francesca Torti, Mauro Pedone, and Edoardo Fibbi. Extreme outliers in surveillance data. *European Commission, Ispra*, JRC132858, 2023.

Marco Riani, Domenico Perrotta, and Francesca Torti. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, 116:17–32, 2012.

Marco Riani, Domenico Perrotta, and Andrea Cerioli. The forward search for very large datasets. *Journal of Statistical Software*, 67:1–20, 2015.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2005.

John Wilder Tukey et al. *Exploratory data analysis*, volume 2. Springer, 1977.