Detecting Extreme Numerical Outliers in Trade Data

A Novel Method for Highly Asymmetric Distributions

Andrea Cerasa, European Commission, Joint Research Centre

UNECE Expert Meeting on Statistical Data Editing, Wien, 7-9 October.

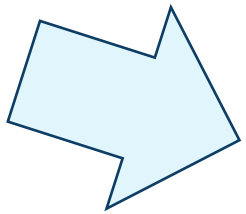
Introduction & Motivation

*The accurate detection of **extreme numerical outliers** in trade data is crucial for effective policy making, anti-fraud measures, and reliable EU-wide statistics.*

Surveillance database of the DG TAXUD is a fundamental tool for facilitating the monitoring of EU trade.

Detailed information of each single import/export customs declaration:

- Product (TARIC classification)
- Gross/net mass
- Economic value
- Origin/destination
- …



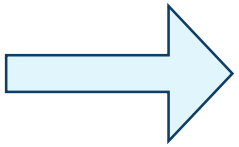
- compiling EU-wide statistics
- ensuring the integrity of supply chains
- combating frauds
- calculating duties

Introduction & Motivation

Huge amount of import/export records registered daily

Almost live transmission of new import/export declarations

Large errors in declared values may occur due to **data quality issues**



These error can severely impact data analyses and lead to incorrect decisions

From Nov-2022 to May-2023, ***1 556 162 485*** Surveillance entries (~1bln/y) regarding ***9 816*** different products

Introduction & Motivation

This motivates our study aimed at defining a method for detecting the extreme values considering the following requirements:

+ ***Flexibility***

The proposed method should be suitable for all the products in the database.

+ ***Statistical properties***

The statistical approach should be easy to apply and to explain, allowing for a strict control of the false alarms.

+ ***Computational efficiency***

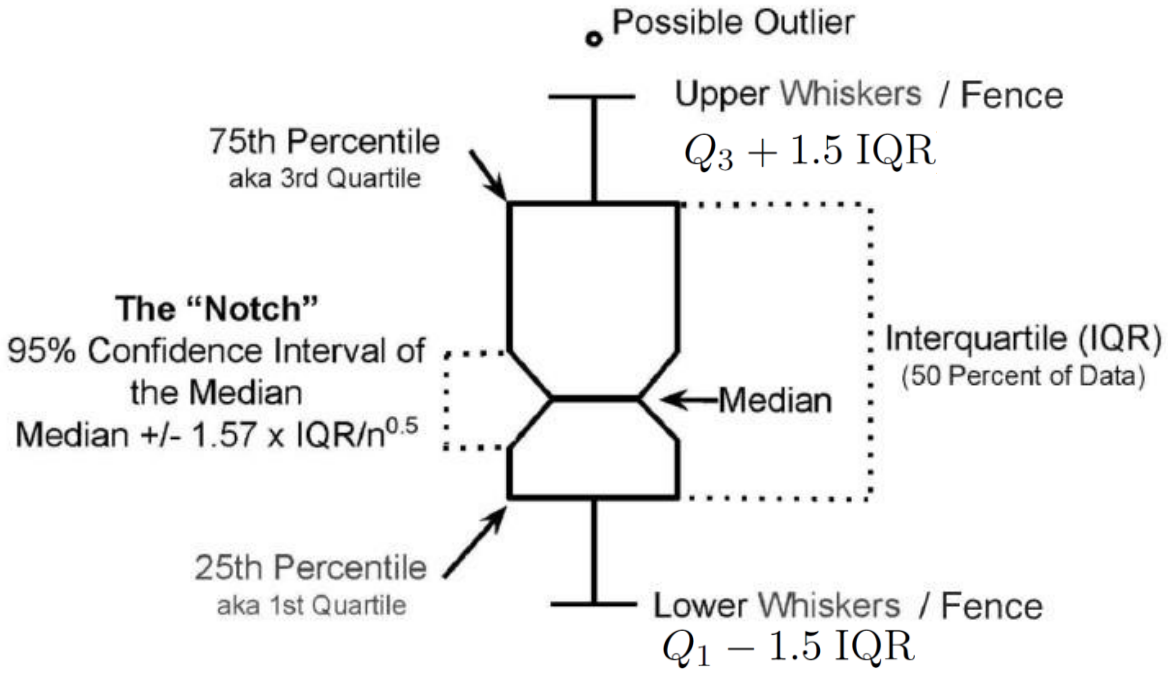
The procedure detects new outliers every day on datasets that could be quite large, and needs to be fast.

+ ***Software simplicity***

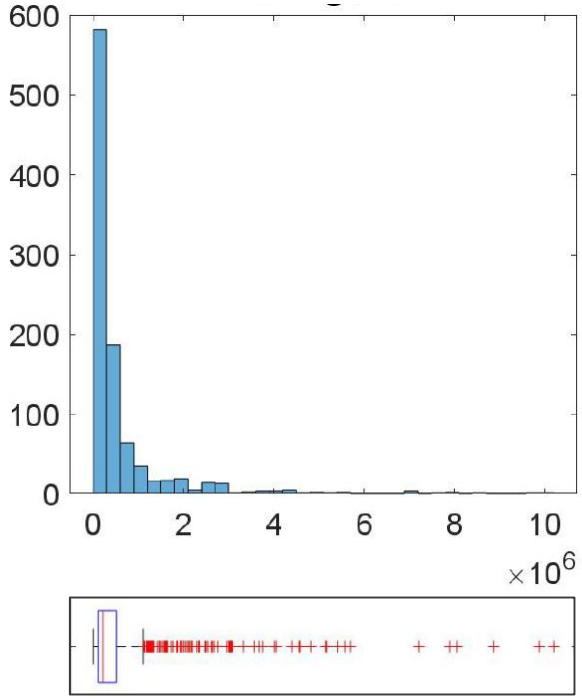
The approach should be based on algorithms that run in an Oracle database with limited calls to specialised statistical functions.

Statistical challenge

One might consider ***classical robust statistical methods*** for detecting extreme outliers in Surveillance data, e.g. the classical **boxplot**.



Boxplot outcomes are not reliable when the distribution is skewed!



European Commission

Statistical challenge

Boxplot adjusted for skewness: new definition of lower/upper whiskers

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

Classical boxplot

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

Adjusted for skewness

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

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

Several specifications for $h_L(\cdot)$ and $h_U(\cdot)$, with $h_L(0) = h_U(0) = 1.5$

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

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

$$a = -3.79$$
$$b = 3.87$$

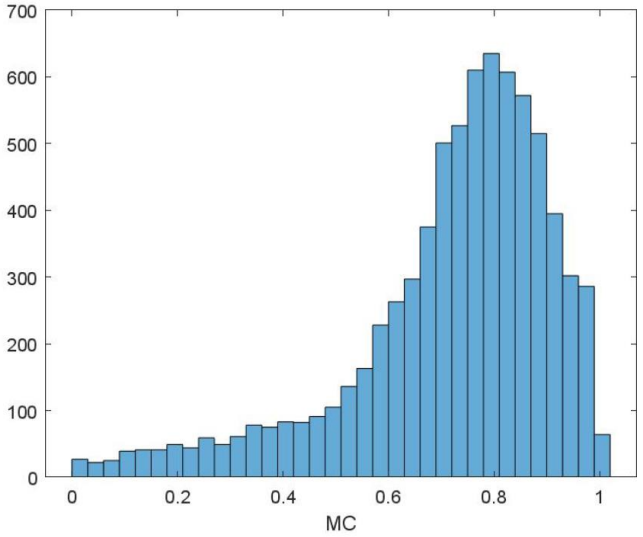
Statistical challenge

But…

The regression for estimating the parameters of the adjusted boxplot relies on simulated distributions with a **moderate degree of asymmetry** ($MC \leq 0.6$).

Surveillance data are characterized by a **higher degree of asymmetry!**

Histogram of the values of the MC calculated on the net mass distributions of the 7 447 products (over 9 816) with more than 100 distinct entries and positive skewness.



Almost 80% of the distributions has a value of the MC larger than 0.6

Our proposal

Considering that:

- ✓ We work with ***real data*** with higher level of skewness than the ***simulated data*** used for deriving the adjusted boxplot;
- ✓ Our goal is ***less ambitious***: we do not aim to extend the classic boxplot to highly asymmetric distributions;
- ✓ we might limit ourselves to consider ***only the right tail*** of the distribution.

Product index
$i = 1, \ldots, 7\,477$

$$\ln(x_{max}^{i} - Q_3^i) \approx \alpha + \beta \cdot \ln(Q_1^i) + \gamma \cdot \ln(Q_2^i) + \theta \cdot \ln(Q_3^i) + \phi \cdot \ln(Q_{95\%}^i) + \lambda \cdot \mathrm{MC}^i$$

Our proposal

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

→ *More flexible specification* with respect to the one used for the adjusted boxplot.

→ The maximum values, quantiles, percentiles, and MCs in the regression come from the *net mass distribution of each product in Surveillance dataset*.

→ By considering $Q_{95\%}$ among the regressors, we are implicitly assuming that *the fraction of outliers in the net mass of each product is less than 5%.*

→ The dependent variable may contain the extreme values we aim to detect. This is why ***a robust method is preferable for estimating the parameters***.

Empirical results – robust estimates

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

	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

Weaker effects for $\ln(Q_2)$, $\ln(Q_3)$ and $\ln(Q_{95\%})$.

Stronger effects for the constant term, $\ln(Q_1)$ and MC.

$\hat{\sigma}$ remarkably higher.

Empirical results – robust estimates

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

	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

Very similar estimates.

They have methodological similarities and use of the same kernel function with settings that ensure an asymptotic efficiency of 95% for both.

Empirical results – robust estimates

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

	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

Stronger effects for $\ln(Q_2)$, $\ln(Q_3)$ and MC.

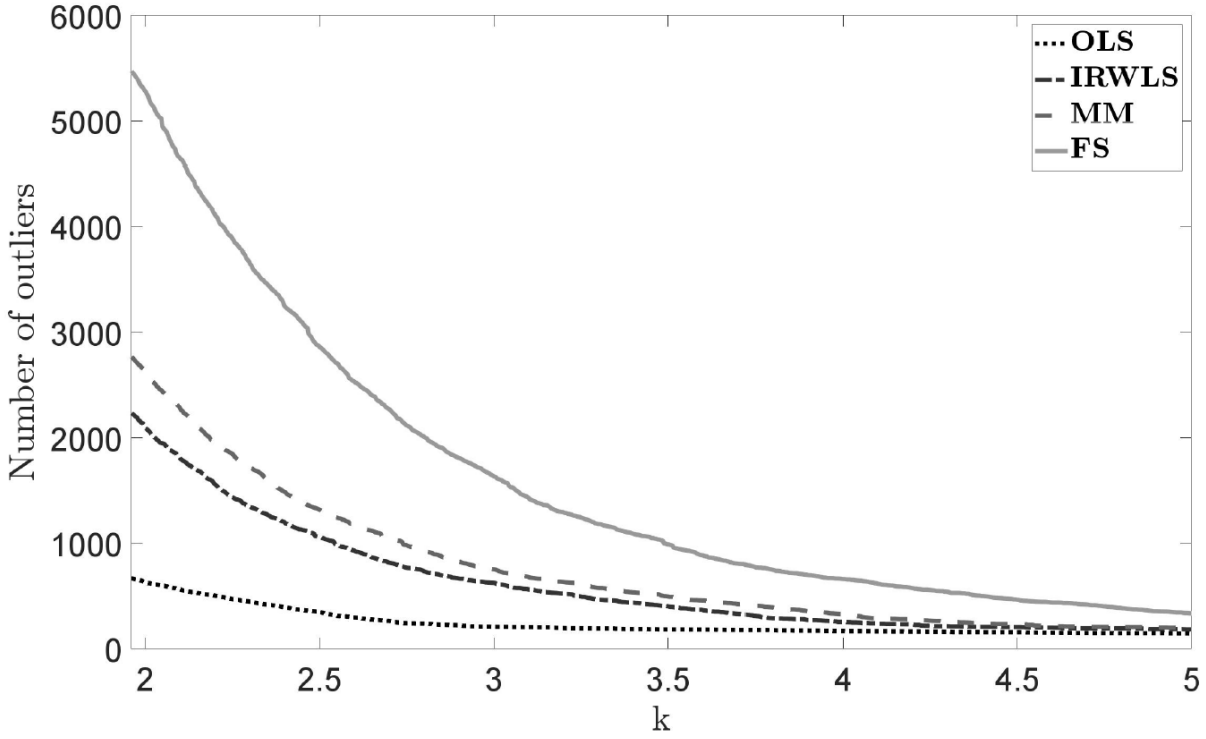
$\hat{\sigma}$ remarkably smaller.

Empirical results – threshold for outliers

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

$$\Downarrow$$

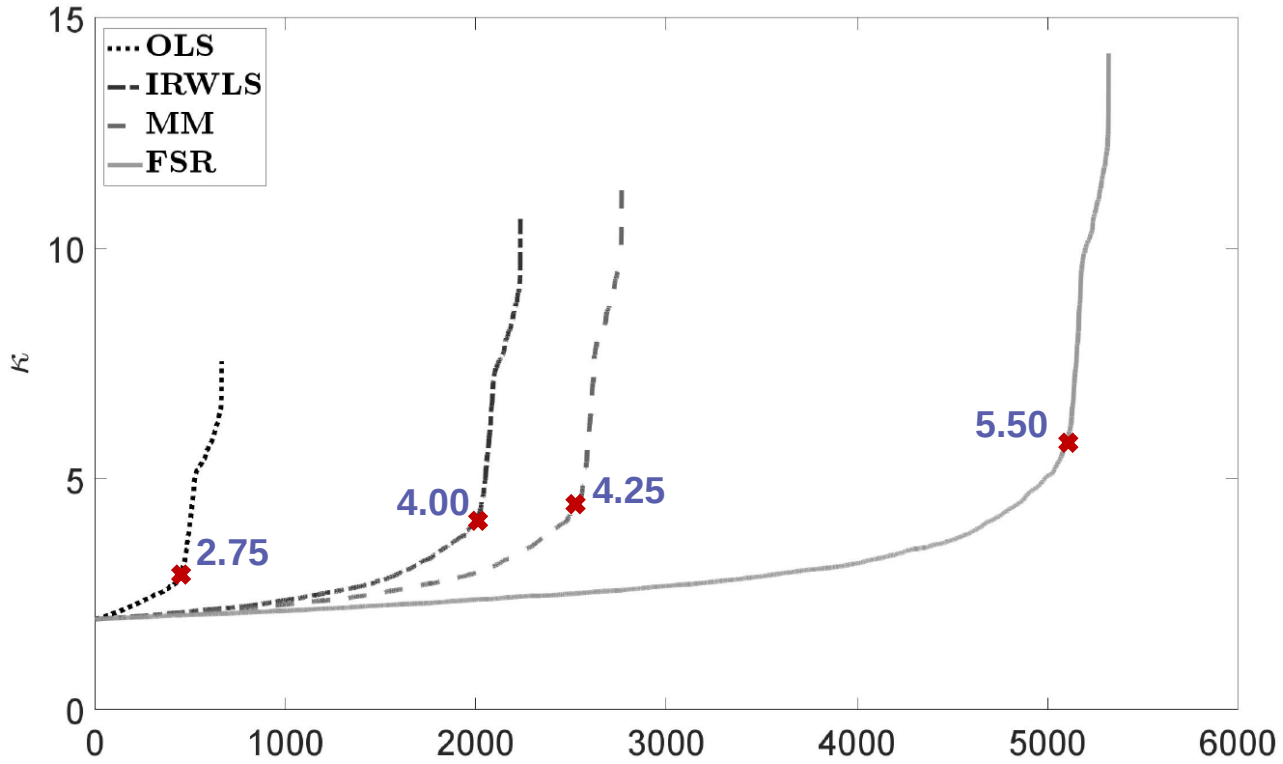
$$x^i_{max} < Q^i_3 + e^{\hat{\alpha}+\hat{\beta}\cdot\ln(Q^i_1)+\hat{\gamma}\cdot\ln(Q^i_2)+\hat{\theta}\cdot\ln(Q^i_3)+\hat{\phi}\cdot\ln(Q^i_{95\%})+\hat{\lambda}\cdot\mathrm{MC}^i+k\hat{\sigma}}$$



European Commission

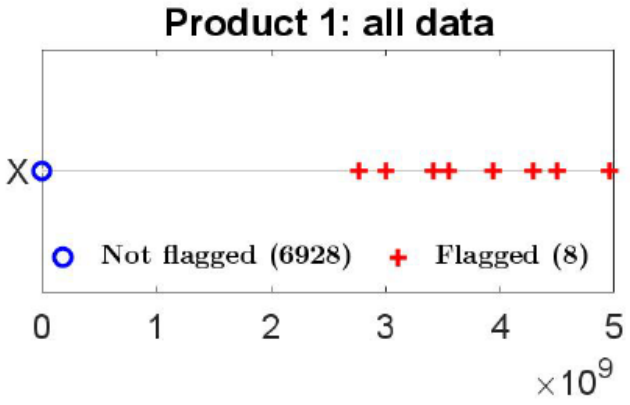
Empirical results – threshold for outliers

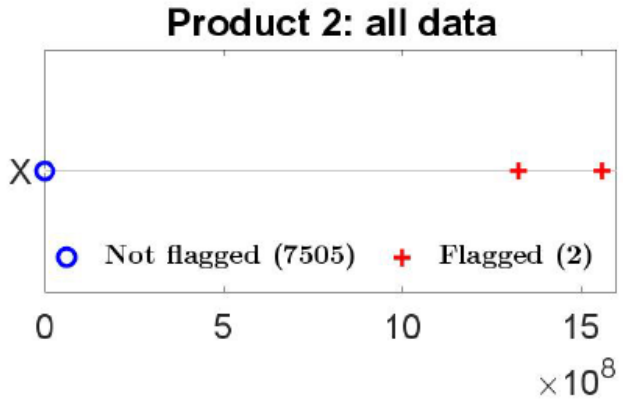
$$\kappa_j^i = \frac{ln(x_j^i - Q_3^i) - \hat{\alpha} - \hat{\beta} \cdot \ln(Q_1^i) - \hat{\gamma} \cdot \ln(Q_2^i) - \hat{\theta} \cdot \ln(Q_3^i) - \hat{\phi} \cdot \ln(Q_{95\%}^i) - \hat{\lambda} \cdot \mathrm{MC}^i}{\hat{\sigma}}$$

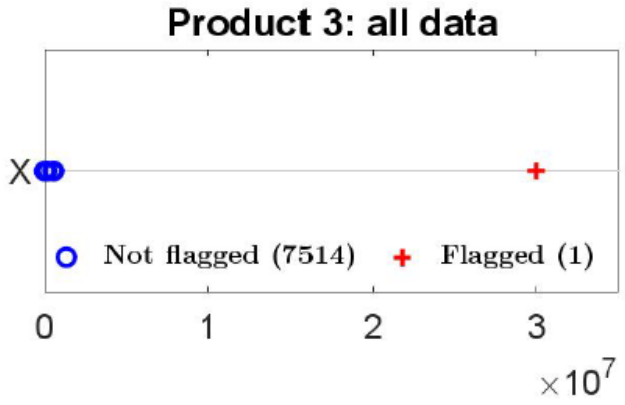


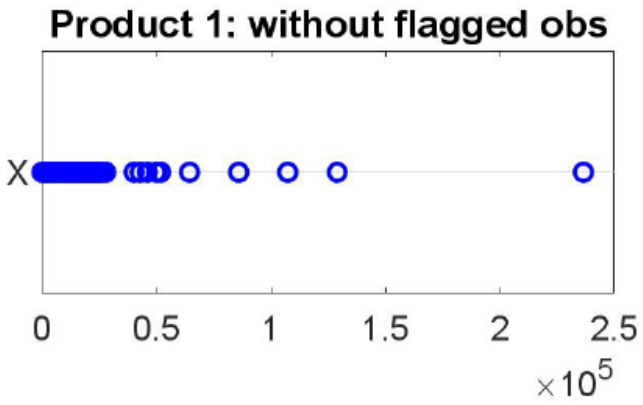
With these thresholds, the number of extreme values flagged by the four methods is approximately 250 (less than 0.00002% of the total number of entries).

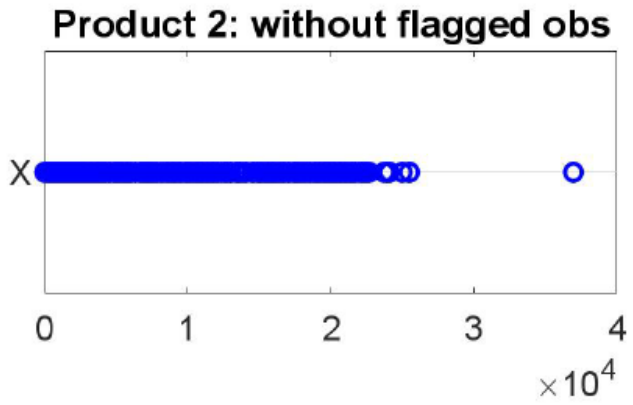
Empirical results – some practical examples

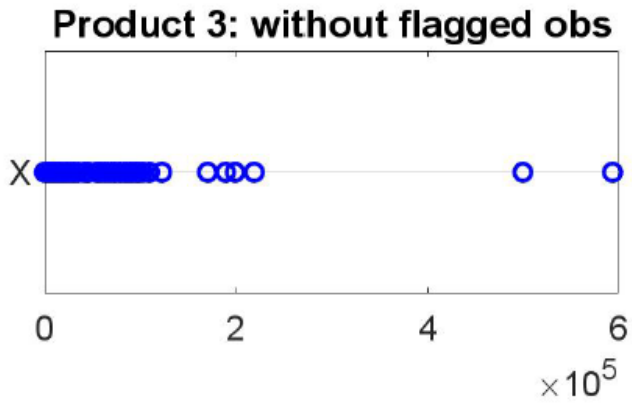












European Commission

Conclusions

Robust statistical approach for flagging extreme numerical outliers suitable for the highly skewed distribution observed in real-world trade data.

The proposed approach meets the specific needs in terms of flexibility, statistical robustness, computational efficiency, and software simplicity.

The empirical application demonstrated its effectiveness in accurately identifying extreme outliers

By ensuring a more reliable outlier detection mechanism, this study contributes significantly to the integrity of international trade data analysis.

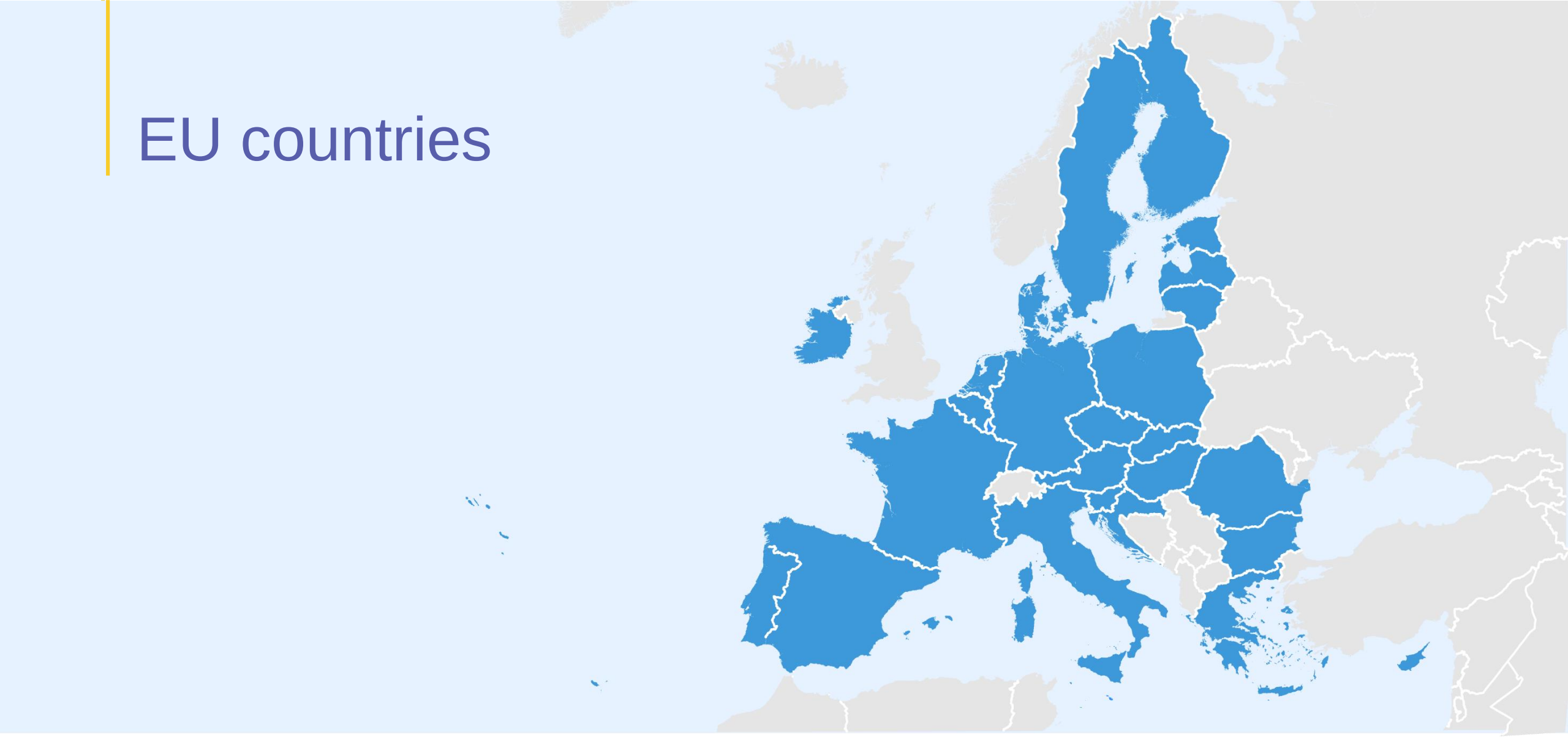
Future work could extend this approach to additional quantitative variables within the Surveillance database

Thank you

EU Science Hub
joint-research-centre.ec.europa.eu

European Commission

EU countries



0 250 500 1,000 Km

European Commission