Statistical Journal of the IAOS 40 (2024) 649–658
DOI 10.3233/SJI-240016
IOS Press

Selective editing for asymmetry analysis in intra-EU trade Micro-Data Exchange (MDE)

Mauro Bruno*, Maria Serena Causo, Giulio Massacci, Francesco Ortame, Giuseppina Ruocco and Simona Toti
Italian National Institute of Statistics (Istituto Nazionale di Statistica – ISTAT), Rome, Italy

Abstract. Since January 2022, the Regulation on European Business Statistics (EU 2019/2152) requires EU Member States to compulsorily share microdata on intra-EU exports. Establishing intra-EU export Micro-Data Exchange (MDE) provides National Statistical Institutes with a new data source to compile intra-EU import statistics. The availability of MDE tackles two key challenges: diminishing the overall response burden on data providers and meeting user expectations regarding the quality of the produced statistics. However, transitioning to a data production system based on MDE data requires the assessment of the coherence and comparability between MDE and National import data.
To identify asymmetries between the two data sources, Istat developed an innovative application designed to foster cooperation among Member States. The tool was developed using the Shiny package in R. The implemented solution allows users to perform exploratory analysis, systematic error detection, and selective editing. The most relevant asymmetries are identified through relative contribution and the asymmetry suspicion indices assessed by *user-defined* thresholds.
Sharing the open tool within the European Statistical System enhances interoperability, promotes method harmonization, and encourages the adoption of official statistical standards.

Keywords: Selective editing, asymmetry analysis, R, statistical standardization

1. Introduction

The availability of the new Micro-Data Exchange (MDE) data source[1] in 2022 has enabled National Statistics Agencies (NSAs) to compile intra-EU imports, reducing the statistical burden on importers. However, before moving from a statistical process based on national Intrastat data collection to using MDE mirror data, it was necessary to detect and assess the cause of the main discrepancies between the two data sources. One of the principles included in the Quality Assurance Framework of the European Statistical System,[2] concerns coherence and comparability, and Intra-EU ITGS[3] data asymmetries raise several concerns about the lack of cross-country comparability. Discrepancies between national and mirror data can have several roots, ranging from the "accidental" presence of outliers to systematic misclassification with respect to products and Partner Countries. In the past few years, National Accounts (NA) and Balance of Payments (BoP) compilers put in place a range of "experimental" statistics in the form of reconciled input-output tables (FIGARO tables,[4] GTAP tables[5]) to address and rectify trade imbalances. There-

*Corresponding author: Mauro Bruno, Istat, Rome, Italy. E-mail: mbruno@istat.it.

[1] With the implementation of the EBS Regulation from 2022 onwards, in addition to data collected or obtained from national sources, NSA have access to intra-Union exports data collected in other Member States (MDE – Micro-Data Exchange), which are exchanged between Member States in a timetable adapted for the production of monthly statistics. A detailed description of MDE implementation can be found at https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Intra-EU_trade_-_exchange_of_micro-data.

[2] https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf.

[3] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:International_trade_in_goods_statistics_(ITGS).

[4] https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210526-1.

[5] https://www.gtap.agecon.purdue.edu/databases/v10/index.aspx.

fore, an MDE-based methodology was developed to detect and address asymmetries, increasing the reliability of the disseminated data.

In this context, the issue of asymmetries in Italian import (IT) data is approached by implementing a tool that exploits MDE data to handle asymmetries, named AsyD. The solution is an application developed in R through the Shiny package.[6] The tool is based on a selective editing approach to detect and isolate the most problematic discrepancies between national and mirror data at a micro-data level. Although the process is still in a preliminary stage, it proved to be very efficient in detecting, assessing, and reconciling asymmetries in 2022 data. The choice of an open-source programming language was driven by the will to encourage cooperation and harmonization of processes across different institutes and agencies [1].

In the following sections, we will describe the context by introducing the data sources and related business problems. Then, we will show the methodology used to detect relevant asymmetries between MDE and national import data. Finally, we will briefly discuss some architectural aspects concerning the integration of AsyD in the production environment.

2. A new European business microdata source: MDE

The new EU Regulation 2019/2152 on European business statistics, which entered into force on January 1, 2022, for international trade in goods statistics, aims at reducing the costs and administrative burden of collecting and producing statistics while improving the quality of the statistical information. It establishes a micro-data exchange (MDE) for intra-Union exports of goods, strongly encouraging innovative and harmonized methodologies for intra-Union trade in goods statistics. The importing EU Member State can use these data in various ways:

- Compile the respective import statistics;
- Amend nationally collected data with missing information;
- Improve data quality and coverage;
- Use the data for analytical work (e.g. for asymmetries resolution);
- Use the data for data analysis and development of new statistical indicators.

In addition to traditional data elements (product, value, quantity, partner country, etc.), the microdata exchange system includes two new key data elements: country of origin and the VAT (Value-Added Tax) identification number of the partner trader in the importing EU Member State. According to Comext June 2023 data, for the year 2022, Member States (MS) estimated 4.052 billion euros for intra-EU total imports, while mirror data estimated 4.109 billion euros. The absolute difference, 57 billion euros, is distributed in different measures for different Member States.

The impact of a single bilateral asymmetry between Member Country A and Member Country B can be measured in terms of contribution *contr (A,B)* to the total asymmetry:

$$contr(A,B) = 100 * \frac{National(A,B) - Mirror(A,B)}{\sum_{i,j}(National(i,j) - Mirror(i,j))} \quad (1)$$

where *National(A,B)* is the nationally estimated import of MS B from MS A, and *Mirror(A,B)* is the export from MS B to MS A, as estimated by MS A.

This analysis aims to develop a tool to highlight and measure such incoherence, facilitating informed decision-making and contributing to the overall reliability and coherence of European business microdata.

3. The implemented solution: AsyD (Asymmetry detection)

The next paragraphs provide an overview of the official statistical standards used as references for implementing the tool. These standards were useful in designing the workflow and developing the main functionalities of AsyD.

3.1. Design principles

The following official statistical standards guided the tool's development for analyzing asymmetries. While the first standard provides guidelines to create and manage a generic statistical data editing process, the other reference frameworks concern service architecture and development. More in detail:

- The **Generic Statistical Business Process Model (GSBPM)**[7] describes and defines the set of busi-

[6] https://www.rstudio.com/products/shiny/.

[7] https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51.

ness processes needed to produce official statistics. It provides a standard framework and harmonized terminology to help statistical organizations modernize their statistical production processes and share methods and components.
- The **Generic Statistical Data Editing Model (GSDEM)**[8] provides standard terminology and models for the development and management of the main data editing functions (Review, Selection, and Treatment). In addition, it describes the relevant metadata monitoring, fostering the automation of the data editing workflow.
- The **Common Statistical Production Architecture (CSPA)**[9] is a reference architecture supporting the standardization of statistical production related to the processes described by the Generic Statistical Business Process Model (GSBPM).

Regarding the GSBPM model, the application implements three specific processes related to *data collection*, *data processing*, and *asymmetries detection* (described in the following section). Based on the GSDEM standard, the activity for asymmetries detection is divided into three phases: *exploratory analysis*, *systematic error detection*, and *selective editing*. This standardized workflow offers a solution that could be reused by different NSAs and organizations, respecting the CSPA principles.

3.2. Methodology

Adopting the frameworks described above, the methods implemented for asymmetries detection are:

- *Exploratory analysis*, to study the distribution patterns of the discrepancies between the two data sources. Starting from the discrepancies calculated in the exploratory analysis, the dataset is divided into four subsets:
  1 Records in which both data sources are available and have the same values by country, product, and operator
  2 Records having a non-zero value in the National data source, and a zero value in the MDE data source
  3 Records having a non-zero value in the MDE data source, and a zero value in the National data source
  4 Records in which both data sources are available and have different values by country, product, and operator.
- *Systematic error detection*, performed on the second and third subsets of records. More in detail, using the non-zero value as a matching variable makes it possible to identify the records affected by misclassification errors in terms of country, product, or both. This approach is applied only in the case of exact matching of the non-zero value in the other source, although corresponding to a different country and/or product. As in this stage, there is no auxiliary information to automatically determine which source is affected by the error, the detected asymmetries need to be investigated manually
- *Selective editing*, based on the relevance of the discrepancies within homogeneous groups or the contribution of each value relative to the total asymmetry. In both cases, the domain expert must choose a threshold to cut off the records according to the resources for interactive editing. Only the records belonging to the fourth subset are eligible for detecting asymmetries through selective editing.

The following section will detail how the selective editing process works, including index calculation and data workflow. It is important to note that the developed solution is limited to the asymmetry identification process aimed at reducing the number of asymmetries to check, while reconciliation has to be manually carried out by the user (such as in [2]).

3.3. Asymmetry detection through selective editing

Selective editing is performed through the following different approaches, relying on the experience of the domain expert, namely:

- The *relative contribution* of each record to the total asymmetry.[10] The percentage of the difference between national data and MDE data, D_i, over the total absolute D, is computed as follows:

$$contr_i = 100 * \frac{D_i}{\sum_i |D_i|} \tag{2}$$

The analysis is based on selecting a threshold for the contribution of each cell to the global asym-

[8] https://unece.org/statistics/documents/2023/01/presentations/generic-statistical-data-editing-models-v-10-0.

[9] https://unece.org/statistics/documents/2013/11/presentations/common-statistical-production-architecture-project-jean.

[10] The asymmetry is defined by considering differences between national and MDE "invoiced values", with no CIF (cost, insurance, and freight) or FOB (free on board) adjustment.

Table 1
Exploratory analysis scheme

Exploratory analysis outcome	Data availability		Value matching	Action
	MDE	National		
Subset 1	Yes	Yes	Yes	No action
Subset 2	Yes	No	No	Systematic error detection
Subset 3	No	Yes	No	Systematic error detection
Subset 4	Yes	Yes	No	Selective editing

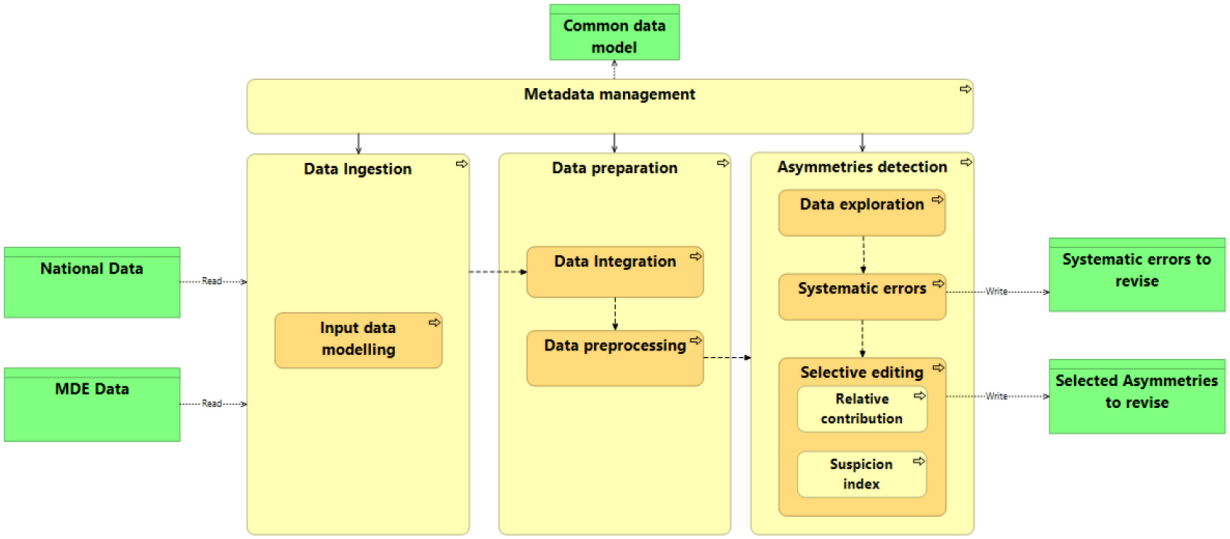

Fig. 1. Data workflow for asymmetry detection.

metry, the contribution of a cell depends on the different grouping options that can be selected by the user.

– A *suspicion index* to evaluate the relevance of the discrepancies within homogeneous groups. The difference between national data and MDE values on a logarithmic scale, $\log R$, is computed on the homogeneous subsets identified by country and four-digit product classification. Then, the first and third quartiles of the empirical distribution of $\log R$, Q_1, and Q_3 are computed. Finally, the suspicion operator [3] is defined as:

$$S_i = \begin{cases} \frac{Q_1 - \log R_i}{Q_3 - Q_1}, & \text{if } \log R_i < Q_1 \\ \frac{\log R_i - Q_3}{Q_3 - Q_1}, & \text{if } \log R_i > Q_3 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Furthermore, we are currently developing various experimental indices, including a *distributional index* that quantifies the proportion of overall suspicion that is not attributable to the observation exhibiting the highest level of suspicion. The primary objective of this index is to distinguish between operators whose asymmetries are uniformly distributed and operators with a higher concentration of problematic observations. We defined this index as:

$$DI = \left(1 - \frac{S_{\text{MAX}}}{S_{\text{TOT}}}\right) \quad (4)$$

3.4. AsyD workflow

The tool implemented the methodology described above through a script developed in R and structured in several steps. This approach was adopted to perform the analysis of asymmetries in an interactive way, taking advantage of the knowledge of domain experts performing data validation. The open-source code can be easily modified to meet specific user needs. The following picture shows the main steps of the data workflow modeled through *ArchiMate* language.[11] The core tasks are grouped into three main sub-processes (Fig. 1):

– Data ingestion: to enable the uploading of data to process and renaming variables according to predefined input data structures for both sources. The accepted input data format is *.csv*, but users

[11] https://www.opengroup.org/archimate-forum/archimate-overview.

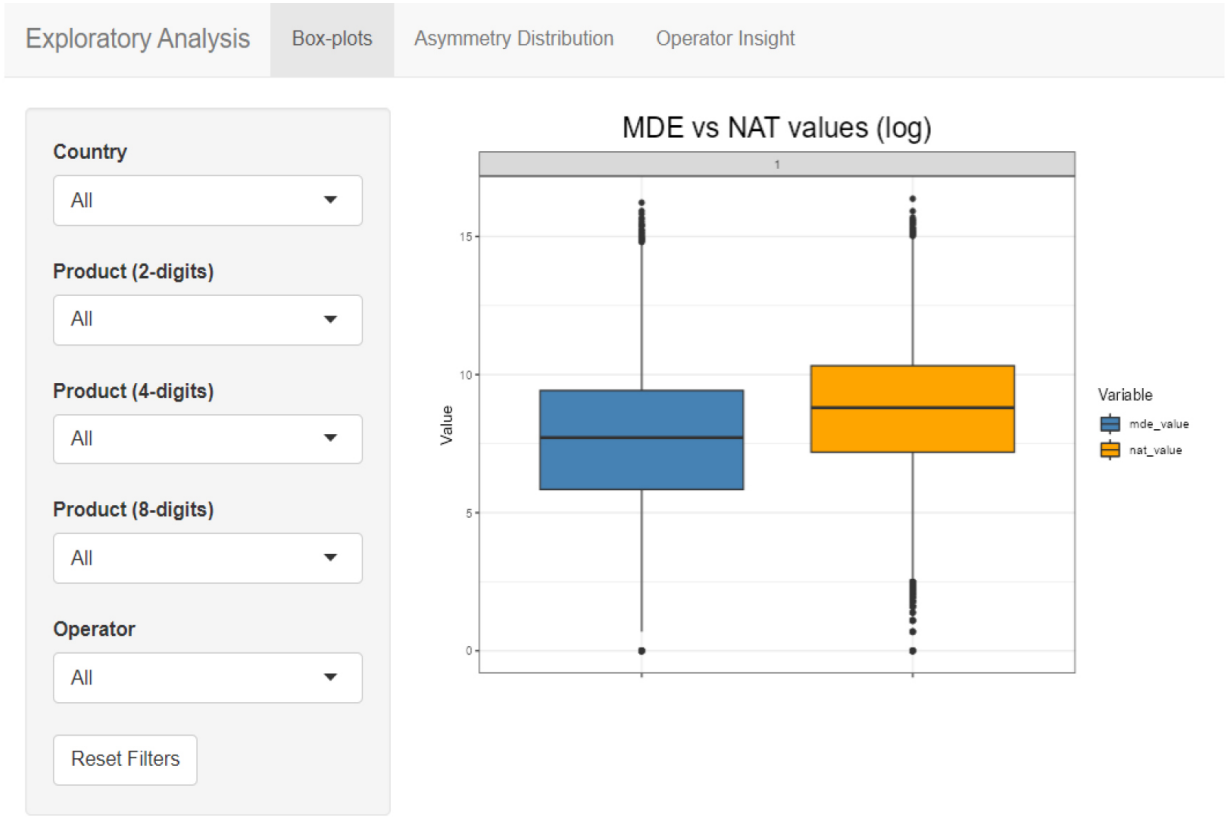

Fig. 2. Exploratory analysis – box plots.

can easily adapt the script to allow compatibility with other data types.

- Data preparation: the sources are linked by country, product, and operator, and the integrated dataset is processed to classify the records in the four subsets previously described, perform the variable transformation, and extract new variables, such as the contribution of each discrepancy to the total asymmetry.
- Asymmetry detection: groups the steps related to the exploratory analysis, the editing of systematic errors, and the selective editing executed through the two approaches described in the previous paragraph. After this process, the domain expert can download a *.csv* file containing the records to check by manual review.

Each task for the asymmetry detection was implemented through a set of modular functionalities, corresponding to a distinct chunk of R Markdown, to provide the user with a clean and easily customizable interface.

Concerning metadata management, the service uses a common data model to standardize input data structures and obtain the input variables to run the code. At the end of the execution, the user may download a list of records for interactive editing. In this case, the output also displays a minimum set of process logs to enable process reproducibility, namely:

- The threshold for the cut-off of the records to revise
- The number of records extracted for asymmetries reconciliation through interactive editing
- The percentage of the records to review out of total records

In relation to the quality assessment, several indicators could be computed after the reconciliation stage to measure the accuracy of the selective editing methods:

- The percentage of errors out of the records resulting from the selective editing
- The comparison between the service output and the results of the procedures executed through other procedures for asymmetry detection.

3.5. *Implemented shiny Apps*

This section describes the different Apps comprising the R Notebook. In particular, the workflow is based on

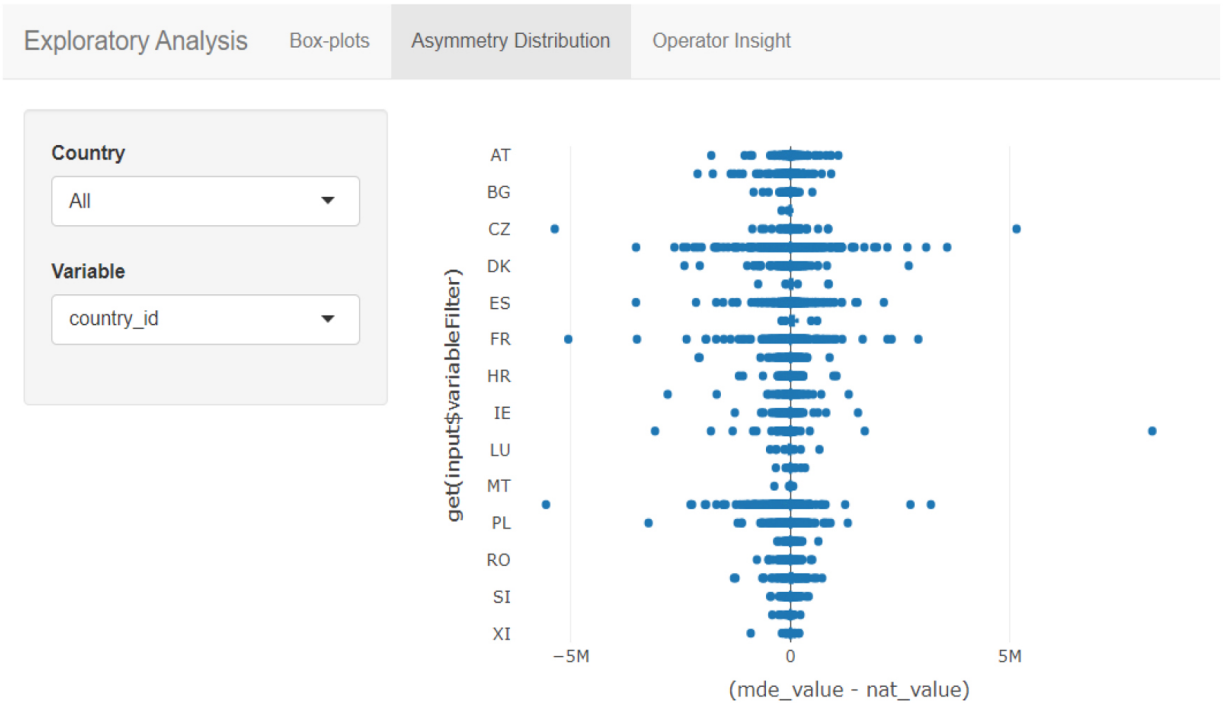

Fig. 3. Exploratory analysis – box plots.

four applications that the domain expert can use to:

1 *Explore* the two datasets, using several interfaces, including box-plots, scatter-plots, and histograms, to conduct early evaluations on data.
2 Observe the *systematic errors* behavior to evaluate their impacts and prioritize the resolution activities.
3 Analyze the *relative contribution*, benefiting from direct real-time calculation and the possibility to export the results.
4 Analyze the *suspicion index*, where results are shown based on thresholds defined via an interface managed by the domain expert.

3.5.1. Exploratory analysis

The "Exploratory Analysis" Shiny App is designed to explore and analyze asymmetry patterns in the data using various visualizations and filtering options. The app allows users to perform exploratory analysis through box plots, asymmetry distribution visuals, and operator-specific measures. Users can dynamically filter the data based on country, product, and operator selections. Figures 2, 3, 4 show the main steps of the exploratory analysis. It was carried out on a sample of anonymized Italian data for the reference year 2022. The *Box-plots* tab (Fig. 2) provides the user with a comparison between the log-distributions of the selected MDE values against the respective National values, representing a first visual step in understanding the nature of the asymmetry.

The *Asymmetry Distribution* tab (Fig. 3) is designed to visualize and analyze asymmetry patterns in the data. Users can explore the distribution of the asymmetry through several filtering options. The *Operator Insight* tab (Fig. 4) is designed to provide users with insights into operator-specific measures within the data. It allows users to filter the dataset by operator and understand their impact on the various asymmetries.

3.5.2. Systematic errors

This tool aims to identify and analyze systematic errors within the dataset by comparing values from the national and MDE data sources for specific operators. Systematic errors occur when the same value for MDE and national data sources are associated with different product codes (Fig. 5).

3.5.3. Selective editing: Relative contribution

This Shiny App offers a deterministic approach to systematically identify and edit data points that contribute to the overall asymmetry. The central concept of this app is the relative contribution to total asymmetry, which is defined as the fraction of the difference

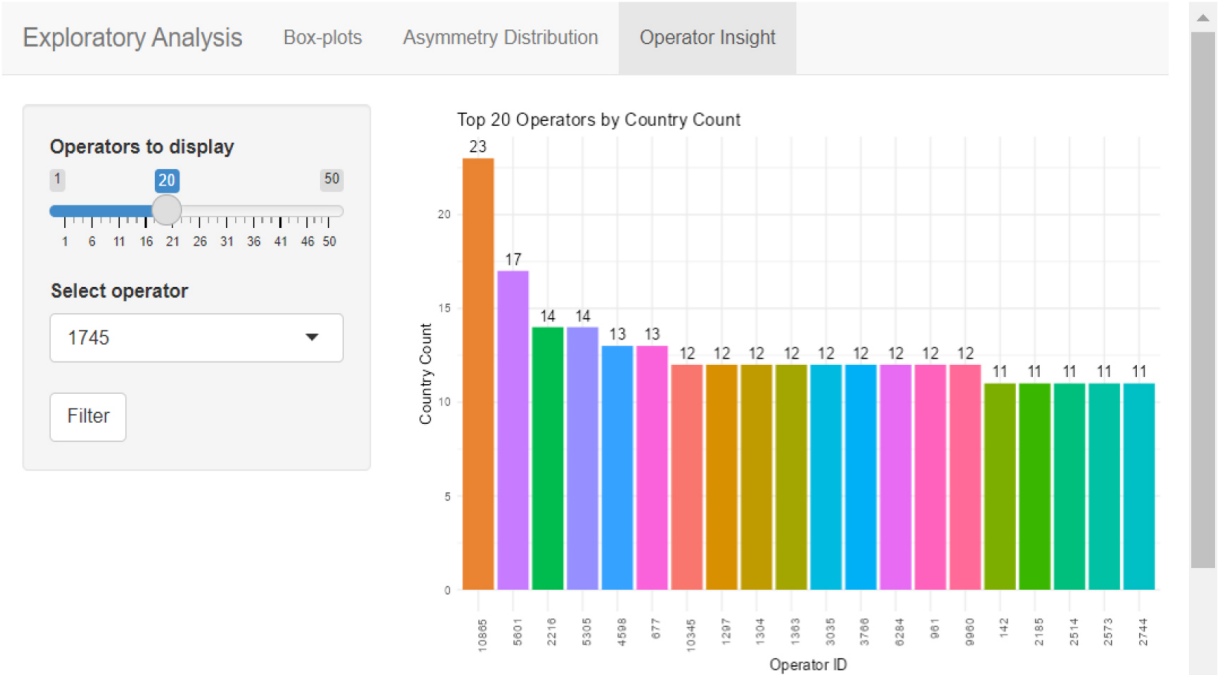

Fig. 4. Exploratory analysis – operator insight.

CSV Search:

	operator_id	value	product_id_8_nat	country_id_nat	product_id_8_mde	country_id_mde
583	5609	504324	1022999	ES	1022951	ES
383	4023	494945	1039190	DE	1039219	DE
200	226	417853	4069073	DE	4069089	DE
214	2357	333838	4069001	LV	4061080	LV
703	6866	306380	1022991	FR	1022949	FR
513	5114	298694	4061050	DE	4069089	DE
172	2066	280512	3074338	ES	3074335	ES
350	367	254014	5051090	DE	5051010	DE
652	6294	245071	1022959	SI	1022999	SI
584	5685	244306	4015039	AT	4041054	AT

Fig. 5. App for systematic errors detection.

between MDE and national values for a specific cell relative to the sum of the absolute differences across all cells. This approach allows for manually identifying significant contributors to the overall data asymmetry. Users can select several grouping columns for data aggregation in any order. In particular, users can choose to group the data by country, product code at different levels of granularity (2-digit, 4-digit, 8-digit), operator, or inclusion status.

Furthermore, users can set a threshold for the absolute contribution (contr) to be considered significant and download the filtered dataset (Fig. 6). Statistical measures such as the relative contribution's mean, median, and standard deviation are shown in the side panel

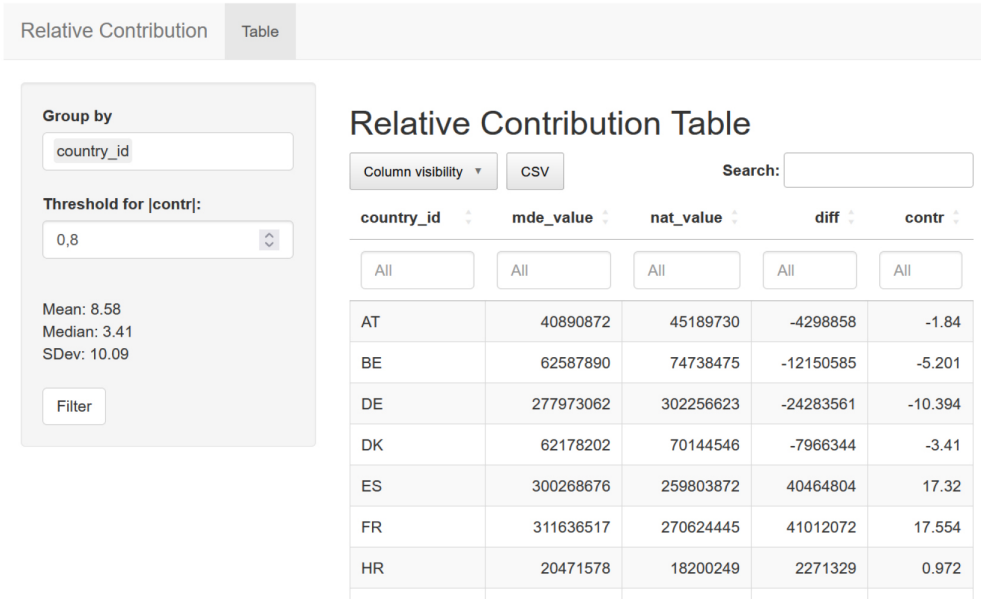

Fig. 6. Shiny App for selective editing through relative contribution.

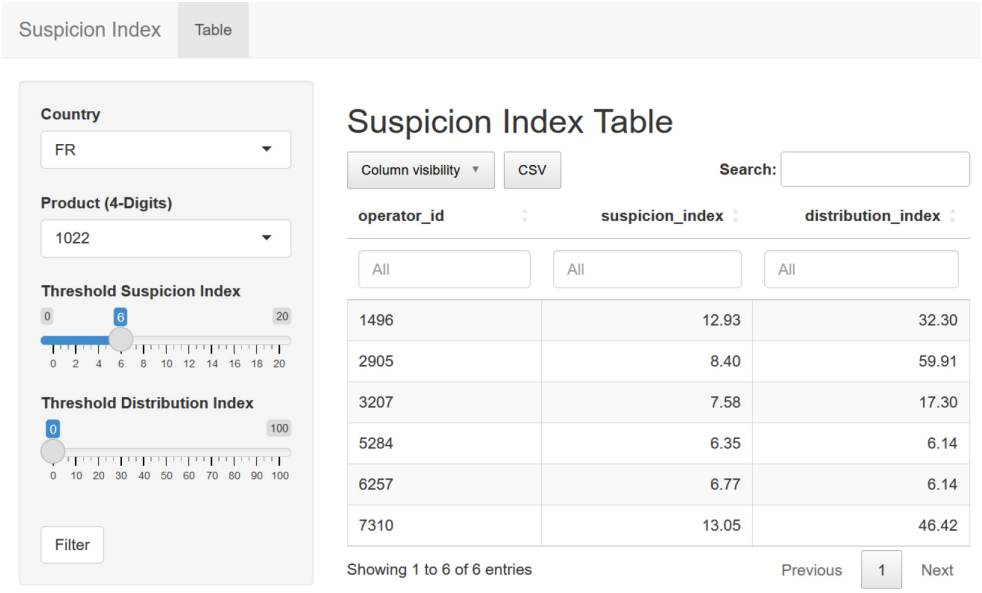

Fig. 7. Shiny App for selective editing through suspicion index.

to guide the user in setting an appropriate value for the threshold. As a reference, in the Italian case, a 1% threshold appeared to be a good compromise.

3.5.4. Selective editing: Suspicion index

This Shiny App offers a systematic approach to identifying specific data points within a dataset. This approach relies on two primary indexes to decide which values to target for editing. The first index (*distribution index*) assesses the distribution of the differences between MDE and National values. The second index, the *suspicion index*, is calculated based on the interquantile range. By combining both indexes, the selective editing approach aims to determine which values in the dataset should be addressed. The suspicion index calculation is a critical aspect of this approach. It as-

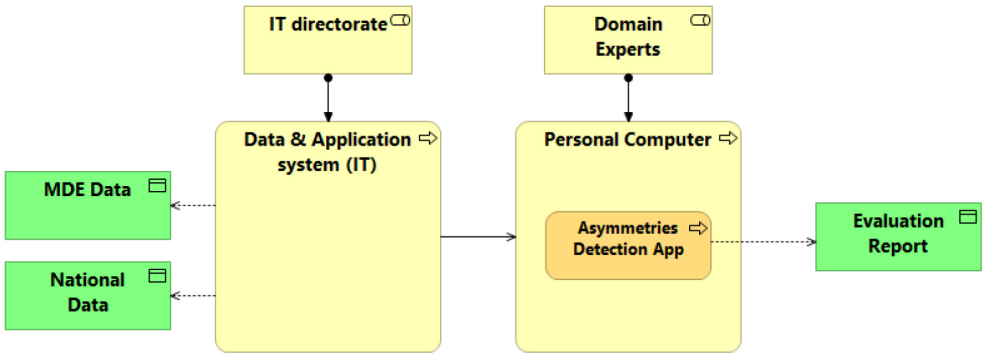

Fig. 8. Local solution design.

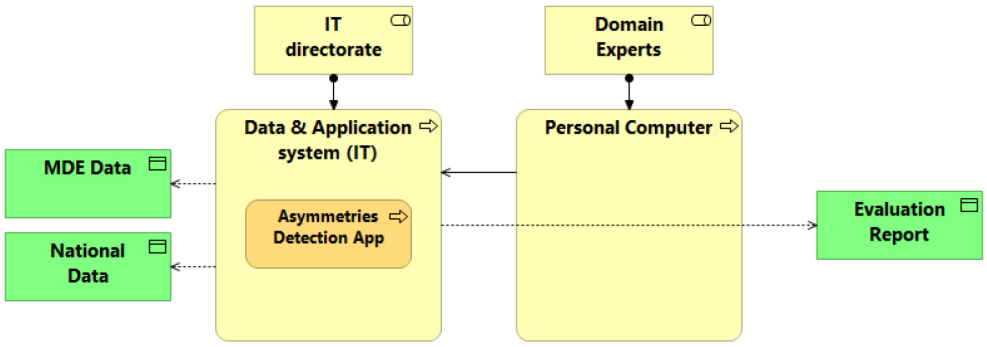

Fig. 9. Remote solution design.

sesses each data point's relationship to the distribution of differences between MDE and NAT values in a log scale. Depending on whether a data point falls within the 25th percentile (Q1) or the 75th percentile (Q3) of this distribution, it is assigned a suspicion value. The distributional index is then calculated based on these suspicion values, providing an overview of the data's contribution to the overall asymmetry.
Users can select the country and 4-digit product code combination they wish to analyze. Additionally, they can set thresholds for both the suspicion and distributional indexes, allowing them to customize their editing criteria (Fig. 7).

3.6. *Expertise requirements and deployment*

The use of AsyD requires minimal IT expertise, comparable to executing an R Notebook. It operates through an intuitive interface, reducing the need for extensive technical knowledge. Users will primarily need to adjust the R Notebook to edit data paths or variable mapping. However, once the application is running, all tasks can be conveniently performed through the user interface, streamlining the user experience and minimizing reliance on coding expertise.

Aligning with the principles outlined in the Quality Assurance Framework of the European Statistical System, it is necessary to understand which deployment solution to adopt, considering several aspects, such as data integration, application usability, quality standards, costs and expertise requirements.

The transition from prototype to a production-ready solution is explored through two main scenarios:

- *Local solution*. Figure 8 shows how the domain expert can run the application on their personal computers, with little to no tweaking. While this option requires no IT expertise, computational limitations arise such as the inability to download full-size CSV files post-revision, a feature that Shiny only allows with server-side processing.
- *Remote solution*. Figure 9 illustrates the deployment of the application on a remote server owned by the Institute, offering benefits such as unrestricted file downloads, enhanced computational resources, and compliance with privacy regulations. However, implementing this solution may necessitate training for IT experts to ensure smooth

operation and maintenance in a production environment.

4. Conclusion

Using the new Micro-Data Exchange (MDE) as a primary data source marks a significant milestone in addressing data asymmetries within intra-EU trade statistics. Implementing MDE, mandated by the EU regulation 2019/2152 on European business statistics, provides Member States with a powerful tool to compile intra-EU exports, reducing the statistical burden on importers. The shift towards MDE-based methodologies is crucial for minimizing the costs and administrative burden associated with data collection while enhancing the quality of statistical production.

The presented workflow, guided by established statistical standards such as GSBPM, GSDEM, and CSPA, underscores the commitment to interoperability, method harmonization, and statistical best practices. The selective editing functionalities, including assessing relative contribution and suspicion index, offer a systematic and flexible approach to identifying and addressing significant contributors to data asymmetry. Some key benefits of the implemented tool include:

- Environment Agnosticism: The application is designed to run seamlessly on various platforms, including web servers, local servers, or local machines. The first two solutions eliminate the need for users to install packages locally, as discussed in [4].
- Integration with R Code: The tool can be integrated smoothly with R scripts already utilized by statistical institutes.
- Enhanced Standardization: The implementation fosters increased standardization among researchers in statistical offices, thereby establishing higher quality standards in the field of official statistics (see [5,6,7,8]).
- Single Framework Solution: The tool allows the construction of an end-to-end solution within a single framework, encompassing both back-end functions and a front-end graphical interface.

The main advantage of the proposed approach with respect to the traditional approach to mirror asymmetries detection[12] consists in the adoption of official statistical standards and in the implementation of a robust partially automated methodology for outlier detection.

AsyD accelerates the analytical process and fosters collaboration and knowledge-sharing among different National Statistical Institutes. The open-source nature of the tools, implemented using R and Shiny, underscores their adaptability and potential integration into diverse statistical production environments with little to no coding expertise. The implemented prototype has received positive feedback from several domain experts, showing that this tool can be effectively integrated into the production environment.

The source code used for this paper is available on GitHub,[13] along with sample data.

References

[1] Vidoni MC. Software Engineering and R Programming: A Call for Research. R J. 2021; 13(2): 600.

[2] Hidiroglou MA, Berthelot JM. Statistical editing and imputation for periodic business surveys. Survey Methodology. 1986; 12(1): 73-83.

[3] Jäder A, Norberg A. A selective editing method considering both suspicion and potential impact, developed and applied to the swedish foreign trade statistics. Macroeconomics and Prices Department, Statistics Sweden; 2006.

[4] Kasprzak P, Mitchell L, Kravchuk O, Timmins A. Six Years of Shiny in Research – Collaborative Development of Web Tools in R. The R Journal. 2021; 12(2): 155-162. Available from: doi: 10.32614/RJ-2021-004.

[5] Choe HM, Kim M, Lee EK. EMSaov: An R Package for the Analysis of Variance with the Expected Mean Squares and its Shiny Application. The R Journal. 2017; 9(1): 252-261. Available from: https//journal.r-project.org/archive/2017/RJ-2017-011/index.html.

[6] MartinkovA P, DrabinovA A. ShinyItemAnalysis for Teaching Psychometrics and to Enforce Routine Analysis of Educational Tests. The R Journal. 2018; 10(2): 503-515. Available from: https//journal.r-project.org/archive/2018/RJ-2018-074/index.html.

[7] Iddi S, Donohue MC. The R Journal: Power and Sample Size for Longitudinal Models in R – The longpower Package and Shiny App. The R Journal. 2022; 14: 264-282. doi: 10.32614/RJ-2022-022.

[8] Weber F, Ickstadt K, and Glass, The R Journal: shinybrms: Fitting Bayesian Regression Models Using a Graphical User Interface for the R Package brms. The R Journal. 2022; 14: 96-120. doi: 10.32614/RJ-2022-027.

[12] https://ec.europa.eu/eurostat/documents/7828051/8076585/Asymmetries_trade_goods.pdf.

[13] https://github.com/istat-methodology/asyd.