European Commission

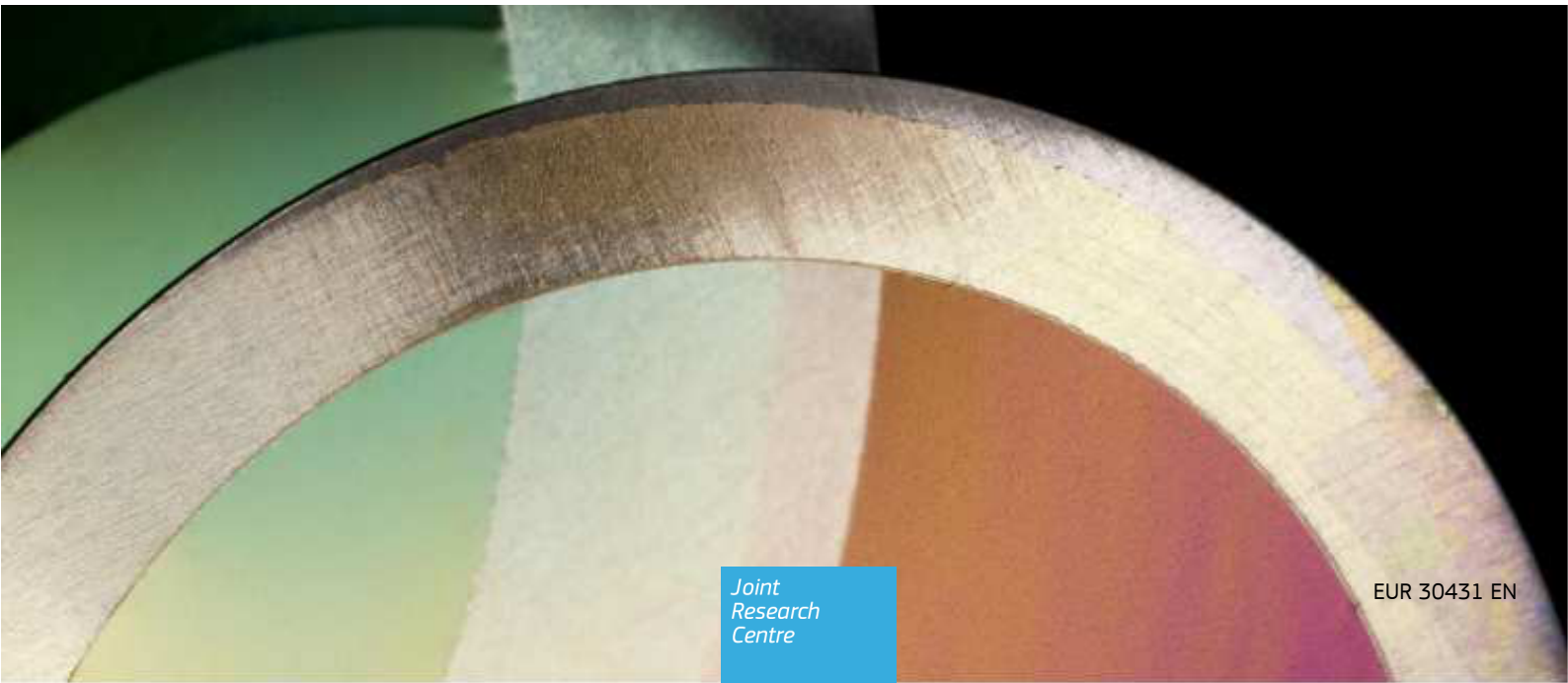
JRC TECHNICAL REPORT

Addressing Price and Weight heterogeneity and Extreme Outliers in Surveillance Data

The Case of Face Masks

Domenico Perrotta, Enrico Checchi, Francesca Torti,
Andrea Cerasa, Xavier Arnes Novau

2020



Joint Research Centre

EUR 30431 EN

Contact information
Name: Francesca Torti
Address: Joint Research Centre, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
Email: francesca.torti@ec.europa.eu
Tel.: +39 0332 786209

EU Science Hub
https://ec.europa.eu/jrc

JRC122315

EUR 30431 EN

PDF ISBN 978-92-76-24707-4 ISSN 1831-9424 doi:10.2760/817681

Luxembourg: Publications Office of the European Union, 2020

How to cite this report: Perrotta, D., Checchi, E., Torti, F., Cerasa, A. and Arnes Novau, X., *Addressing Price and Weight heterogeneity and Extreme Outliers in Surveillance Data - The Case of Face Masks*, EUR 30431 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-24707-4, doi:10.2760/817681, JRC122315

This page is left intentionally empty

Contents

About the authors

Two longstanding parallel JRC areas of expertise found common points of interest in this work.

— Domenico Perrotta, Francesca Torti, Andrea Cerasa and Enrico Checchi (JRC.I.3, Text and Data Mining Unit) conduct research and development activities in statistics and information technologies for anti-fraud, trade and security. The statistical contributions discussed in the report have been developed over the years in collaboration with their academic partners, acknowledged below.

— Xavier Arnes Novau (JRC.G.II.7, Nuclear Security Unit) works in the field of dual use and export control. He is a customs expert and executes research in the relevant fields, together with the colleagues acknowledged below.

Acknowledgements

Cristina Versino and Filippo Sevini (JRC.G.II.7, Nuclear Security Unit) have benefited this joint activity with their in-depth knowledge of international trade data, and enhanced the results obtained with a marked data visualisation and navigation mindset. We also thank Dimitra Triantafyllidou (DG TAXUD) for the careful reading of the manuscript, her suggestions on delicate subject matters and for urging reflection on subtle technical issues (in particular, the effect of our sampling approaches).

The deployment of the IT tools is by Giuseppe Sgarlatta, Emmanuele Sordini, Marzia Grasso and Patrizia Calcaterra of the 'Text and Data Mining Unit' of the JRC. The efficient exploitation of Surveillance data was made possible by a large-scale database infrastructure designed by Aris Tsois and maintained by Mauro Pedone, of the same Unit.

Statistical methods and models have been developed with the substantial contribution of academics from the University of Parma, primarily Prof. Marco Riani and Andrea Cerioli. Thanks also go to Prof. Mia Hubert and Peter Rousseeuw (both at KU Leuven), for their key role in the development of the robust time series model used in our monitoring system, and to the authors of the TCLUST methodology: Agustin Mayo Iscar, Luis Angel García-Escudero and Alfonso Gordaliza (University of Valladolid).

The problems related to the COVID-19 case study could be readily addressed in the middle of the pandemic thanks to software resources available in the FSDA toolbox for MATLAB, which we developed jointly with the University of Parma over several years. FSDA is in Github (`https://github.com/UniprJRC/FSDA`), in the File Exchange of Mathworks (`https://it.mathworks.com/matlabcentral/fileexchange/72999-fsda`) and its documentation is accessible at the address `http://rosa.unipr.it/fsda.html`. JRC has also ported some key FSDA tools to the SAS and R environments, with significant contributions from Dr Aldo Corbellini (University of Parma) and Dr Valentin Todorov (United Nations Industrial Development Organization, UNIDO).

This wide line of work is financially supported by the institutional budget of the JRC and two Administrative Arrangements. The main one, with OLAF under the Hercule III Programme, aims at the development of Automated Monitoring Tool's (AMT) for its investigation units and their partners in the Member States. The second, with DG TAXUD (S-DAC), provides scientific support for the development of new financial and security risk criteria for the Member States Customs. JRC thanks their colleagues in these services for the continuous trust and support given to this research.

Abstract

The Customs Surveillance system (Surveillance Monitoring System (SMS) or Surveillance) of DG Taxation and Customs Union (DG TAXUD) centralises all European Union import and export declarations, collected from the national customs authorities on a daily basis according to Article 55(2) of Commission Implementing Regulation (EU)2015/2447. The Customs Surveillance data provide actual and prompt information about quantities, values, origin and destination of each traded commodity. The analysis of these trade flows poses a number of statistical challenges caused by the heterogeneous nature of the trade and the presence of many anomalous numerical values in the declarations (clerical errors, market peculiarities but also frauds). This report presents some solutions to these data analysis complications.

The statistical approaches discussed in the report have been jointly developed by the Joint Research Centre and its academic partners, and have been applied in various broad application areas using our FSDA software library, based on the MATLAB environment. Their illustration is driven here by the specific needs of the Clearing House Task Force established by the European Commission to monitor the trade of COVID-19 related commodities during the pandemic which exploded in 2020. We show how this activity contributed to a relevant policy impact, by supporting the work for refining the definition of the codes used to trade protective face masks, adopted in the EU as of October 2020.

In addition, the report is addressed to the services of the European Commission that are responsible for the Customs Surveillance system (DG TAXUD) and for its use in anti-fraud (European Anti-Fraud Office, OLAF). The Member States Customs authorities also benefit from these studies, because our methods and models are deployed in a customs anti-fraud resource jointly developed and maintained by the JRC and OLAF in the respective IT environments.

1 Motivation and problem description

The European Commission can monitor EU trade in great detail thanks to the *Surveillance* database of DG TAXUD. This is a customs "surveillance" system that collects on a daily basis from the national authorities all EU trade data, as ruled by EEC (1993) and Article 55(2) of EEC (2015).

In Surveillance, each import/export transaction contains information recorded by the trade operators in a customs declaration, including in particular weight, quantity or supplementary units, value, origin and destination of the consignment (the form used for the declaration is called Single Administrative Document, SAD). The product type is specified according to a hierarchical coding system specified inside the TARIC database, which includes the wide customs tariff of the Union, also managed by DG TAXUD. For illustration, Table 1 reports the taxonomic ordering of a specific animal product and one from textiles; note the different level of detail in the description of the final TARIC level codes.

The Surveillance and TARIC systems allow monitoring of EU trade for different purposes: policy making, EU-wide statistics, securing the supply chains, anti-fraud, and also facilitating the economic operators in their activities, for example determining licensing requirements or duties to be paid. Small differences in codes can make big differences in duties and licenses, and sometimes the temptation to miss-declare the code or origin of a product is strong. For all these reasons, TARIC is updated on a daily basis by the European Commission, with the aim of reacting to those issues that might arise or evolve with regard to: preferential tariff rates, tariff suspensions, third country duties or tariff quotas. Sometimes, new codes are created from scratch to account for completely new products; more often the existing classification is refined by changing description and splitting or merging codes as in the examples of Figures 1 and 2.

Unlike TARIC, the Annex I of the Common Customs Tariff (Combined Nomenclature) is amended annually, under the supervision of the Committee on Tariff and Statistical Nomenclature. The Harmonised System is even more stable and robust since it is amended every 5 years by the World Customs Organization, the current version in force being HS-2017, valid until the end of 2021. However, given that the Harmonised System works at 6-digit level, it is less accurate at describing goods than TARIC.

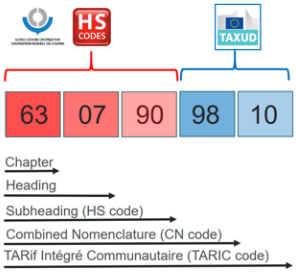
In such a rapidly evolving world it is often impossible to adopt the necessary classification refinements in due time. For example, goods like face masks that during the 2020 COVID-19 emergency have become suddenly critical, were classified together with other heterogeneous products under the same codes (see Section 2). The procedure initiated to introduce more specific codes needs several iterations between the services involved in the revision (the requesting service, Eurostat, DG TRADE, DG TAXUD and in some cases the World Customs Organization, WCO) and therefore time to be completed.

Given that the revision process of TARIC is long and it is not applicable to all needs, one should be able to analyse Surveillance data and other trade sources using methods capable of highlighting possible

63	Other Made-Up Textile Articles; Sets; Worn Clothing and Worn Textile Articles; Rags	**HS** Chapter
6307	- Other made-up articles, including dress patterns:	**HS** Heading
6307 90	- - Other	**HS** Subheading
6307 90 98	- - - Other	**CN** code
6307 90 98 10	- - - - Nonwovens	**TARIC** code

02	Meat and edible meat offal	**HS** Chapter
0202	- Meat of bovine animals, frozen	**HS** Heading
0202 20	- - Other cuts with bones in:	**HS** Subheading
0202 20 50	- - - Unseparated or separated hindquarters:	**CN** code
0202 20 50 11	- - - - of bison	**TARIC** code
0202 20 50 15	- - - - other	**TARIC** code

Table 1: The EU classification system in TARIC. It consists of three main hierarchical components, the Harmonized System (HS) that is unique at international level, the Combined Nomenclature valid in the EU, and the additional TARIC code that determines the applicable duty rates and other customs measures (see figure on the right). The table shows the taxonomy of a specific product in the 'Textiles and textile articles' section (that covers HS Chapters from 50 to 63) and 'Live Animals; Animal Products' section (that covers HS Chapters from 01 to 05). Even at the ten digits of the TARIC, commodities can have different levels of specification.



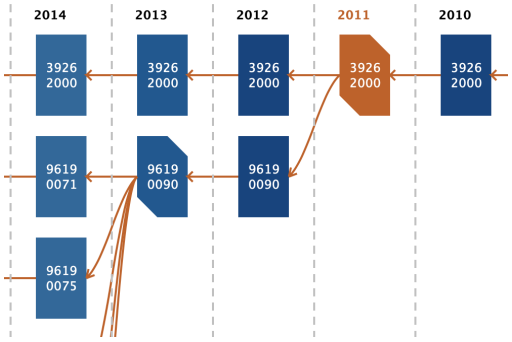


Figure 1: Split of TARIC code 39262000 in 2011 (*Articles of apparel and clothing accessories produced by the stitching or sticking together of plastic sheeting, incl. gloves, mittens and mitts, excl. goods of 9619*).

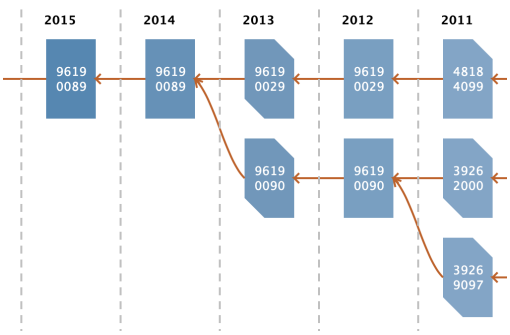


Figure 2: Merges leading to TARIC code 96190089 in 2014 (*Sanitary articles, e.g. incontinence care articles, excl. of textile materials, and sanitary towels, tampons, napkins and napkin liners for babies*).

Source of Figures 1 and 2: HERMES section of the THESEUS web resource of the JRC, which is based on TARIC data. The section is publicly available at `https://theseus.jrc.ec.europa.eu/index.php?id=1600`.

data sub-groups and revealing relevant price or weight-per-unit heterogeneity([1]). The information on the fine-grained data structure allows the uncovering of the real trade volumes and price markets of the critical commodities, and also points to attempts to evade duties or circumvent customs measures.

The JRC uses consolidated statistical instruments to identify anomalies when products are rather homogeneous, in the sense that data are formed by a dominant population possibly affected by a certain amount of contamination. For example, in Perrotta and Torti (2010) we approached the problem of detecting price outliers in regression on monthly aggregates of traded values and quantities, in Riani et al. (2018) we addressed a related price estimation problem complicated by potential small sample size issues, while in Rousseeuw et al. (2019) we also considered the detection of anomalies in time series of such trade flows.

To address product heterogeneity, that is to identify several homogeneous sub-products with large heterogeneity among them, it is natural to use clustering methods, but it is important to be aware of several key problems and know how to solve them properly. The main issue concerns the automatic choice of the number of potential groups (sub-products) in the data and also other model-specific parameters that too often even the specialised literature sets with tacit assumptions or leaves at the margins of the discussion.

In this report, we illustrate our approaches to the heterogeneity problem with the case study introduced in Section 2. Section 3 introduces an approach that we experimented with during the COVID-19 emergency, which analyses the distribution of the unit prices (the values divided by the quantities) using a *non-parametric kernel method* tailored to the specificity of trade data. Then, Section 4 illustrates the application of a consolidated *clusterwise linear regression* approach that we developed in a robust setting to account for the presence of anomalies (errors, market peculiarities, fraudulent activities) and concentrated noise (the so called 'small trade area', formed by a large number of unimportant trade transactions).

The two approaches are complementary and have different merits. The latter is along consolidated parametric models that we studied extensively in recent years, which seem to capture well the structure of trade data and that we adopted already in operational tools for anti-fraud (Arsenis et al., 2015; Perrotta et al., 2020). The former does not make assumptions about the distribution of the unit prices, is computationally simple and easily interpretable by non-specialists. The evaluation of the relative performances of the two approaches on Surveillance data is left to a separate study.

A last non-trivial problem that this report addresses is the detection of outrageous anomalies, such as clerical errors or mistranscribed digits that appear in data as very extreme outliers in quantity, values and (when applicable) supplementary units. Section 5 explains why these types of errors are difficult to detect precisely without over-declaring the potential anomalies, and illustrates a solution that the JRC has implemented in the Surveillance monitoring system in THESEUS. Section 6 introduces a complementary activity, aimed at monitoring if the Surveillance data flows properly from the Member States to the database of DG TAXUD. Conclusions and next steps are in Section 7, and are followed by the technical appendices. These last ones are not meant to be comprehensive, but formalise the statistical problems and contain links to the relevant literature.

([1]) Also data with some level of aggregation, such as COMEXT, contain cases of heterogeneity that can similarly be addressed.

2 Use case: analysis of import prices of face masks

Filtering face-piece (FFP) masks for respiratory protection (FFPs, N95, N99, etc.), single-use disposable masks, 'community masks', surgical masks, dual face masks (universal masks), have become a very common commodity during the COVID-19 crisis, especially for the emerging shortages at certain time frames. Monitoring their import/export has become a policy priority, in order to optimize availability in healthcare settings in a context where supply might be very limited. Therefore, trading these products under correct classification is of paramount importance. Unfortunately, currently the relevant categories available in the Combined Nomenclature and TARIC are rather broad.

Normally, the proper respiratory protection, Filter Face Piece (FFP) masks fall under CN code 6307.90.98, but other products can be traded under the same code, including surgical masks and also other types of goods such as umbrellas, surgical drapes, decorative textile articles, cushions, covers for cars and baskets for cats. Fortunately, TARIC splits the CN code 6307.90.98 into three more precise ten-digits codes, as shown in Table 2, but this refinement does not solve all the problems:

— Code 6307.90.98.10 should cover most of the protective masks non-compliant with safety standards in the market, including 'community masks' and other textile masks which are knitted or crocheted. This is the closest to the technical description of the protective masks and is probably the description most used by economic operators during the COVID crisis. For this reason, it should be the main code to monitor.

— Code 6307.90.98.91 is more generally for non-woven and hand-made textile articles.

— A residual code 6307.90.98.99 is used for those textile articles that do not meet any characteristic or description throughout the whole of chapter 63. It is likely that some protective masks are declared by economic operators also under this residual code probably due to their lack of knowledge of interpretation of Tariff classification rules, and therefore this code should be monitored closely too.

The last code in Table 2 (6307.90.10) that refers to textile masks knitted or crocheted, may also be used for declaring 'community masks', but also completely different products such as fans, bags and eye masks for sleeping. Finally, note that operators may also use code 4818.50.00 for mixing masks that should fall under code 6307.90.98.10 with unrelated products like cellulose paper masks.

It is obvious that this articulated classification system complicates the possibility to monitor precisely the trade of the protective masks or other specific products of interest, especially in residual codes like 6307.90.98.99. One way to distinguish the different products is to consider that they can differ in both price and specific weight; for example, the grammage for the FFP masks is at least $200\,g/m^2$ while the surgical masks have a lower specific weight. Statistical methods can be used to identify precisely the fine-grained structure of the import/export declarations.

The analyses that follow were done within this framework, in support to the decisions of the *COVID-19 Clearing House for medical equipment*, which is operating in the Secretariat General to facilitate the timely availability in the EU of the medical supplies needed to fight the virus. The results of this and other types of analyses are made available by the JRC to the Clearing House, the Commission Services and the Customs

Category	Product description	TARIC code	TARIC description
Face and eye protection	Textile facemasks, without a replaceable filter or mechanical parts, including surgical masks and disposable facemasks made of non-woven textiles. This includes the masks known as N95 Particulate Respirators. Note: the heading also includes N95 respirators with simple exhalation valves as these remain respirator masks and are not gas masks.	6307.90.98.10	Other made-up articles, including dress patterns – Nonwovens
		6307.90.98.91	Other made-up articles, including dress patterns – Hand-made
		6307.90.98.99	Other made-up articles, including dress patterns – Other
		6307.90.10.00	Other made-up articles, including dress patterns: knitted or crocheted

Table 2: Commodities with TARIC codes including face masks. The broad TARIC descriptions indicate that these codes are used to classify also other products. Facemasks mostly fall in codes 6307.90.98.10 and 6307.90.98.99. The Category and Product descriptions are derived from the WCO and EU standards and other information sources, and are those that we use in the THESEUS monitoring system of the JRC.

in Member States through two monitoring systems, the THESEUS web resource and a TABLEAU-based data visualisation system (Figures 3 and 4 show the entry page of the COVID-19 sections of these systems; Figure 5 is a view on a specific product category). The JRC Technical Report by Arnés-Novau et al. (2020) describes this activity in great detail, with focus on the technical aspects of the products monitored.

Two sub-periods are taken into account: the first runs from January 2019 to February 2020, in order to build a 'pre-COVID' benchmark; the second runs from March 2020 to May 2020, including the most critical part of the emergency. The comparison of the two distributions allows the assessment of the effects of COVID on face masks import prices. We carried out two complementary analyses:

1. Price approach: this consists of analysing the unitary price, that is the price paid for one Kg. More precisely, the distribution of the unit prices logarithm is estimated through a non-parametric procedure based on the Kernel Density Estimation described in Section 3.

2. Value-weight approach: this consists of a cluster-wise regression analysis of the traded value (dependent variable) and weight (independent variable). In this approach we identify groups with state-of-the-art methods that are very flexible and *robust* to the presence of outliers. The main method used is called TCLUST-REG (García-Escudero et al., 2010b), which has been studied and experimented with several years in the context of international trade data (see for example Torti et al., 2018).

The case study required two additional ingredients, generally needed when analysing Surveillance data.

A. In some cases the number of declarations to analyse can be very large even if we focus on one day only (tens of thousands of records). This makes estimations challenging, both computationally and statistically, especially in the 'value-weight' analysis. In such cases we apply an additional sampling step that reduces the number of points without losing the structure of the informative part of the data (in Cerioli and Perrotta, 2014, we described the problem and proposed a first solution in relation to TCLUST-REG). The approach is explained in Section B.3.

B. Our statistical methods are robust to the presence of outliers: therefore, our estimates remain stable even if a large proportion of the data is anomalous. However, our monitoring systems include graphical views on the data and summary statistics which rely on the source Surveillance data; therefore, a single clerical error can completely distort these statistics. These far outliers are detected and filtered out prior to any other analysis as illustrated in Section 5.

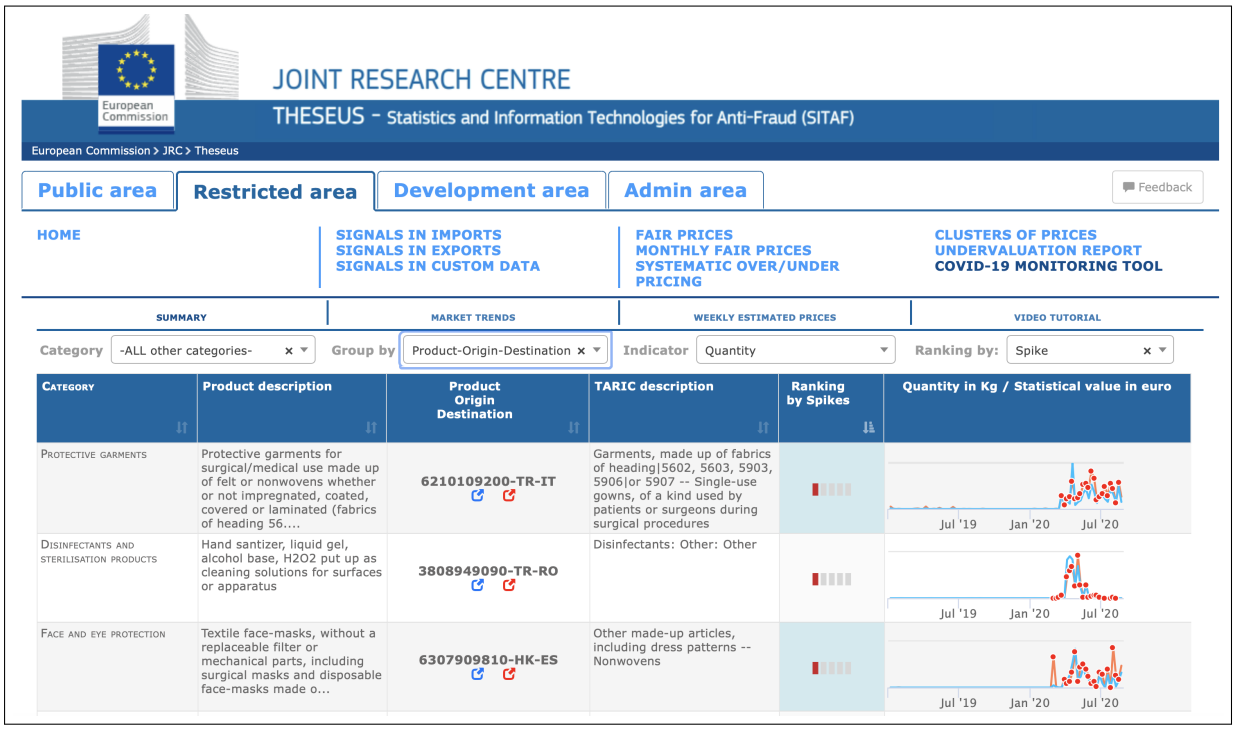


Figure 3: The entry page of the COVID-19 monitoring tool of `https://jrc.theseus.jrc.ec.europa.eu`, a web resource of the JRC mainly designed for users in anti-fraud services.

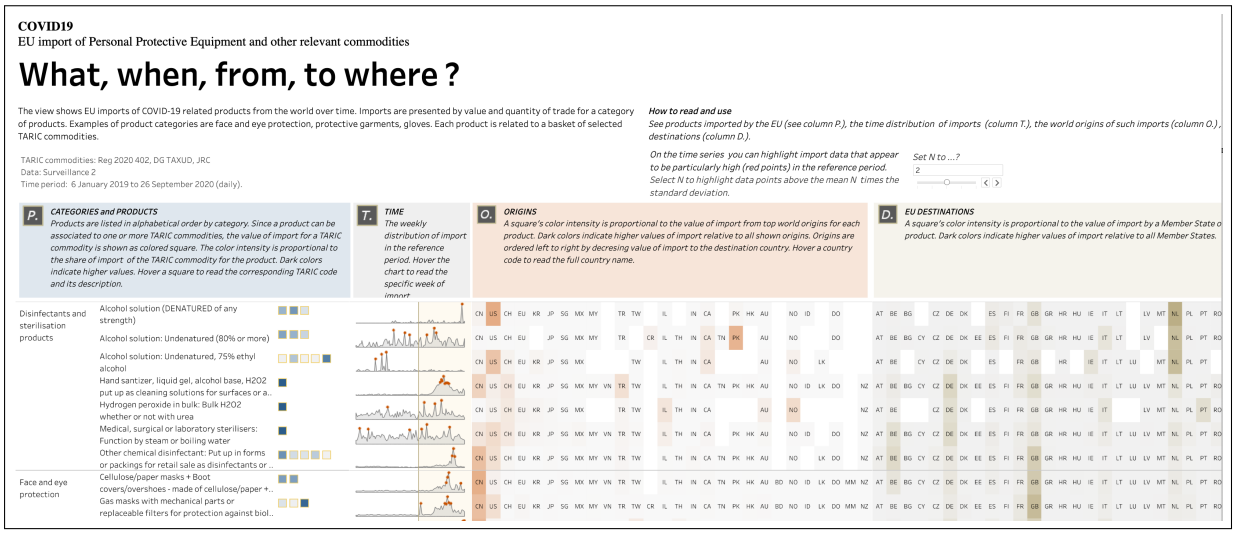


Figure 4: The entry page of the COVID-19 dashboards, which are accessible from `https://visualise.jrc.ec.europa.eu/`, a TABLEAU server of the JRC.

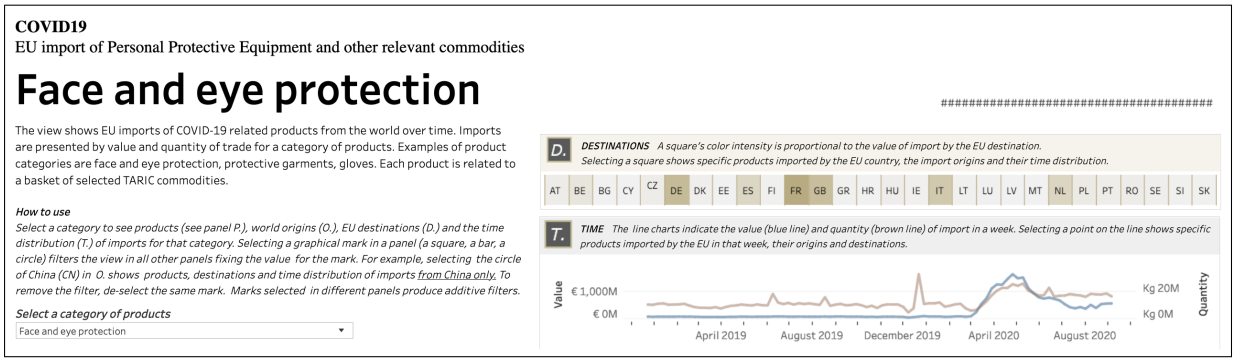


Figure 5: A COVID-19 dashboard on "face and eye protection". The dashboards are interactive and provide views by Product, Origin, Destination and Time; this figure shows the last two views.

3 Application of kernel density estimation to unit prices of face masks

We computed the unitary prices, that is the prices paid for one Kg of face masks and we estimated the distribution of the logarithm of the weighted unit prices. Section A formalises the approach.

3.1 The market before and during the pandemic

Here we consider two subsequent sub-periods: the first runs from January 2019 to February 2020 and can be seen as a 'pre-COVID' benchmark; the second runs from March 2020 to May 2020 covering a critical part of the COVID crisis. *We expect to be able to assess the effects of the COVID crisis on the face masks import prices from the comparison of the two distributions.*

Figure 6 presents the results obtained in the benchmark period (pre-COVID). The estimated density distribution (left panel) highlights two modal import prices centered around 2.3 and 6.9 euro/kg respectively, but the corresponding bells overlap considerably. This suggests the presence of a sort of continuum of prices between the two modes. This is in fact confirmed by the scatter plot in the central panel, where a clear discrimination between the two prices does not emerge. The scatter plot in the right panel, where the transactions associated to the four codes are highlighted, shows that the scatters of four codes are quite homogeneous in terms of net weight, statistical value and unit price.

Figure 7 presents the results of the same analysis for the COVID period (March 2020 – May 2020). The plots provide a completely different picture and clearly highlights the shock of COVID on the face masks market. A new modal import price appears (left panel) clearly higher than that in the previous period. The separation between low-price and high-price masks is even more evident in the two scatters (central and right panel), where the points are clustered in two well separate groups. The scatters also show how the higher-priced masks represented the majority of the imports.

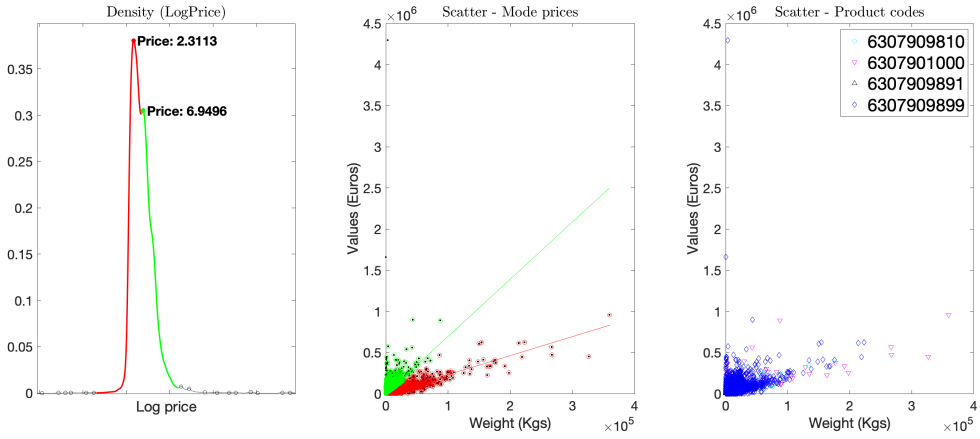


Figure 6: KDE price analysis of products in Table 2. Pre-COVID period. KDE suggests two main market prices, but not well separated.

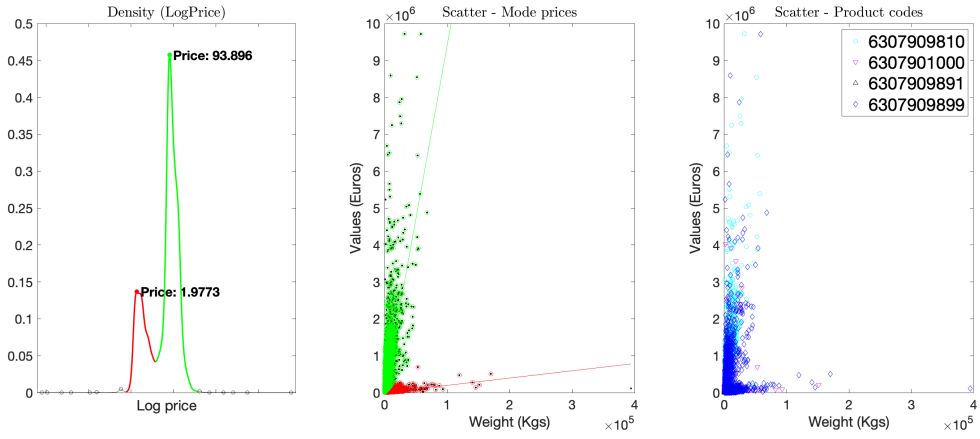


Figure 7: KDE price analysis of products in Table 2. COVID period. There are two well separated market prices.

Product code	Low-Price face masks (Import Price per Kilo < 16.61) No of imports	Weight (kgs)	Value (euros)	High-Price face masks (Import Price per Kilo ≥ 16.61) No of imports	Weight (kgs)	Value (euros)
6307901000	6,677	7,126,103	29,326,408	14,496	2,163,500	162,593,825
6307909810	9,342	6,129,367	34,613,129	131,406	54,579,361	6,434,620,437
6307909891	152	169,465	750,197	533	173,817	14,208,833
6307909899	92,406	47,979,558	204,035,688	292,724	34,096,572	3,456,959,775
Total	108,577	61,404,493	268,725,422	439,159	91,013,250	10,068,382,870

Table 3: Low and High Price face masks imports from March 2020 to May 2020.

		Total	Monthly average
Pre-COVID Period (Jan 2019 - Feb 2020)	**No of imports**	887,677	63,046
	Weight (kgs)	422,319,166	30,165,655
	Value (euros)	2,325,334,139	166,095,296
COVID Period (Mar 2020 - May 2020)	**No of imports**	547,736	182,579
	Weight (kgs)	152,417,743	50,805,914
	Value (euros)	10,337,108,292	3,445,702,764

Table 4: Assessment of COVID effect on the market volumes of the four face masks product codes.

Table 3 documents the magnitude of this pattern. For each code, the number of imports involving high-priced face masks is at least more than twice than the number of imports of low-priced face masks ([2]). The separation is even more notable if we look at the values involved, that for the high-priced masks is almost 40 times higher (€269 million versus €10,068 million). Moreover, it is interesting to note that the codes mainly used for importing the high-priced masks were 6307.90.98.10 and 6307.90.98.99.

The COVID emergency influenced not only the price distribution of the four product codes used for importing face masks, but also their market volumes. From the figures presented in Table 4, it is possible to quantify its effect: on average, the number of imports of face masks products almost tripled, the weight almost doubled, whereas the value involved is more than 20 times higher.

Finally, Figure 8 focuses on the main commodity that should include also protective masks, that is code 6307.90.98.10. In this case, the pre-COVID period confirms that the market of the commodities in this specific code was rather homogeneous, with a dominant estimated price of 3.66 euro/kg. Then, during the COVID crisis, the share of the market with this price reduced considerably, while two unprecedented ultra-expensive markets (accounting for 90.35 and 264.42 euro/kg respectively) emerged stoutly.

As a final remark, it is worth stressing that it is not possible to determine with high precision which share of these codes is actually related to face masks. Analysts of the relevant services believe that the low value products represent other goods, which were the major goods traded under these codes. After March, trade in all goods of Chapter 63 not related to COVID-19 has collapsed, these goods included. It is possible that only the two high price clusters are actually related to face masks. Note that as of October 3rd, the classification has changed and masks will have their own codes: therefore, as soon as more data will become available, it will be possible to see whether this is the case.

3.2 Monitoring the market over time

Now, instead of considering two non-overlapping subsequent periods, we take a moving window of 1 month starting on January 2019, and we estimate the sequence of KDE prices. For each date, we report the probability density estimate in a surface plot showing in hotter colours the prices that are more likely to be observed. Figure 9 shows a clear discontinuity between March and April 2020, which clearly corresponds to the insurgence of the COVID crisis. This graphical tool can be used for a rapid screening of the more critical commodities. The extension of the tool for inferential purposes is under study. Several application examples are discussed in Arnés-Novau et al. (2020).

([2]) The threshold on the unit price (16.61 €/kg) has been chosen on the basis of the unit price distribution for the period March 2020 - May 2020 (left panel of Figure 7). Precisely, the threshold is the point in the middle of the two bells, corresponding to the lower price (in green and red) and the high one.

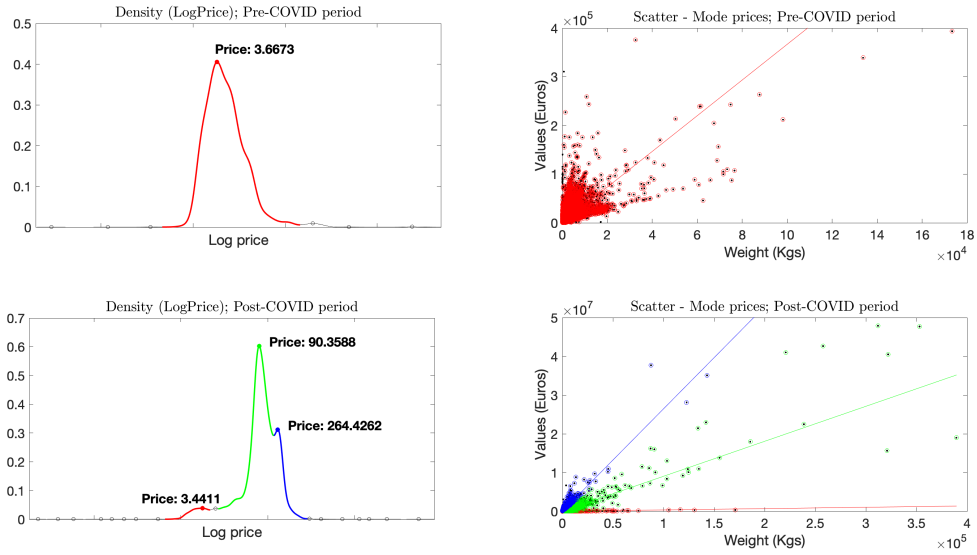


Figure 8: KDE price analysis of commodity code 6307.90.98.10 of Table 2 before (upper panel) and during (bottom panels) the COVID crisis.

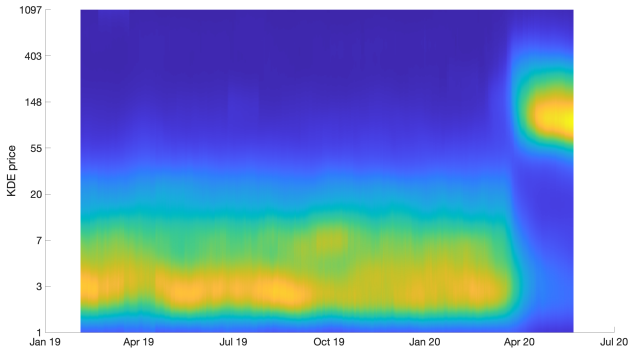
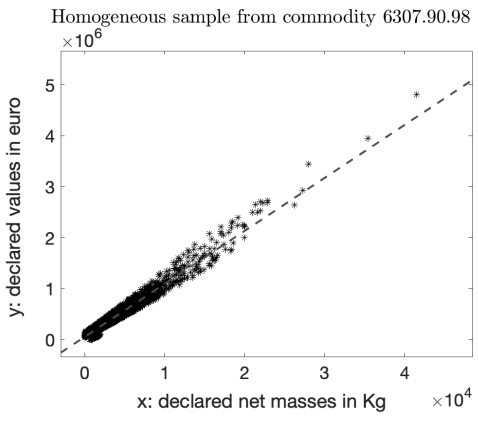


Figure 9: KDE price analysis of commodity code 6307.90.98.10 of Table 2, in a moving window of 30 days. Hot colours indicate a higher probability of observing a given price. The jump at the end of March is striking.

4 Application of robust cluster-wise linear regression to values and weights of face masks

In this section, we analyse the data from a perspective that differs from the approach in section 3.

We start again from n records of the Surveillance database (y_i, x_i), with y representing the declared (statistical or customs) values and x_i the corresponding quantities. When data refer to comparable transactions, for a same homogeneous commodity, we can reasonably assume that the declared values are proportional to the traded quantities through a *constant commodity price*, that is we have a linear relation between y (the response/dependent variable) and x (the explanatory/independent variable) forced to go through the origin, so that zero quantity necessarily implies zero value. The figure on the right illustrates the typical linear structure between value and quantity declarations; the slope of the linear fit (dashed line) is an estimate of the price, which is about 100 €/kg in this specific data sample.



A similar reasoning can be done when y represents the quantity and x the number of supplementary

units imported, and the two variables are proportional via a *constant weight per unit*. Although the above advocates for a lineal model without constant term (the intercept), we know that data may contain records with zero quantities and positive values and vice versa (these cases concern records with values or quantities lower than certain thresholds, coming e.g. from e-commerce([3])). For this reason, our treatment can follow a conventional linear model with intercept, which has the advantage of avoiding controversial issues that the regression without intercept poses (see for example Eisenhauer, 2003).

Now, our motivating case is that even for specified TARIC code, trading partners and transport means, often the data to analyse appear heterogeneous. This means that data do not follow the nice single linear structure in the preceding figure; instead, they cluster around G homogeneous subsets, each one generated by a distinct linear model. An appropriate framework to address these data is the *cluster-wise linear regression*, described with some mathematical formality in Annex B. This section concentrates on the results that can be obtained with this approach on the face masks case study.

To identify the distinct homogeneous groups in the dataset, we use the TCLUST-REG method (Section B.2) that is capable of leaving a fraction α of most outlying points unassigned and at the same time constrains the relative dispersion of the groups below a threshold c, to avoid detecting spurious structures formed by observations that incidentally are almost perfectly aligned. In order to estimate α and c and understand how many groups G form the dataset, we use a tool that applies TCLUSTreg on the dataset several times, for different $\{G, \alpha, c\}$ combinations, and monitors a statistic that can spot the best model combination (details and references are in Section B.4).

In the COVID period considered (March - May 2020) there are 515,089 records, which are too many to apply a sophisticated model-based method like TCLUST-REG without incurring in complications of both a statistical and computational nature. In order to reduce complexity and to avoid serious estimation issues, we applied two subsequent sampling steps. The first consists in selecting a random sample of 30,000 observations weighted by the norm of the observations, which is the Euclidean distance of each observation from the origin of the axes; in doing so, we have avoided to select the declarations of negligible value and/or quantity. The second sampling step consists in selecting from the 30,000 observations a reduced subset that preserves the cluster structure but at the same time avoids keeping too many uninformative observations that concentrate in particular areas of the scatterplot, typically where the values and quantities declared at the Customs are both small (Section B.3 formalises the small trade area'). This second step reduces the sample to less than 1,000 observations, without losing information on the overall data structure.

Figure 10 illustrates the results obtained on a final sample of 779 observations that can be found in the FSDA clustering datasets folder, under the name `facemasks.mat`. The top-left panel shows the monitoring of TCLUST-REG for various percentages of trimming α and number of groups G (to simplify discussion, we fixed the restriction factor $c = 64$, which is a value that leaves a lot of flexibility to TCLUST). The plot reports in the y axis a model selection criterion that should be minimum when G is optimal. Some curves reach the minimum at $G = 5$ for all α values but for $\alpha = 1\%$ the optimal option is $G = 3$. The top-right panel shows another instrument that monitors also the stability of the optimal solutions. Roughly speaking, the longer and thicker the rectangle is for a certain number of groups, the better. From this perspective, $G = 3$ and $G = 2$ are the most stable across the trimming levels which are monitored, and $G = 5$ is confirmed to be the best for $4‰ \leq \alpha \leq 6‰$. The scatter plots at the bottom show the classifications of the observations for $G = 2$ and $G = 3$ when $\alpha = 6\%$, with the corresponding estimated prices reported in the caption of the Figure: the prices are consistent with those estimated with the kernel approach of Section 3.

Figure 11 shows that results are more uncertain if we monitored TCLUST-REG on the larger sample of 30,000 observations selected only on the basis of the trade declaration size. In fact, both graphical instruments seem to indicate that the number of potential groups could have still increased.

These results can be replicated using functions in the Flexible Statistics for Data Analysis (FSDA) toolbox available as 'Add-On' inside MATLAB or in github (`https://github.com/UniprJRC/FSDA`). The documentation can be found at `http://rosa.unipr.it/FSDA.html`. We report our codes in Annex D. The specific functions used for this purpose are also mentioned in the technical annexes of the report.

([3]) For other trade statistics, like COMEXT, the problem is even more pervasive, as a result of higher statistical thresholds fixed by Member States within the limits permitted by Community legislation.

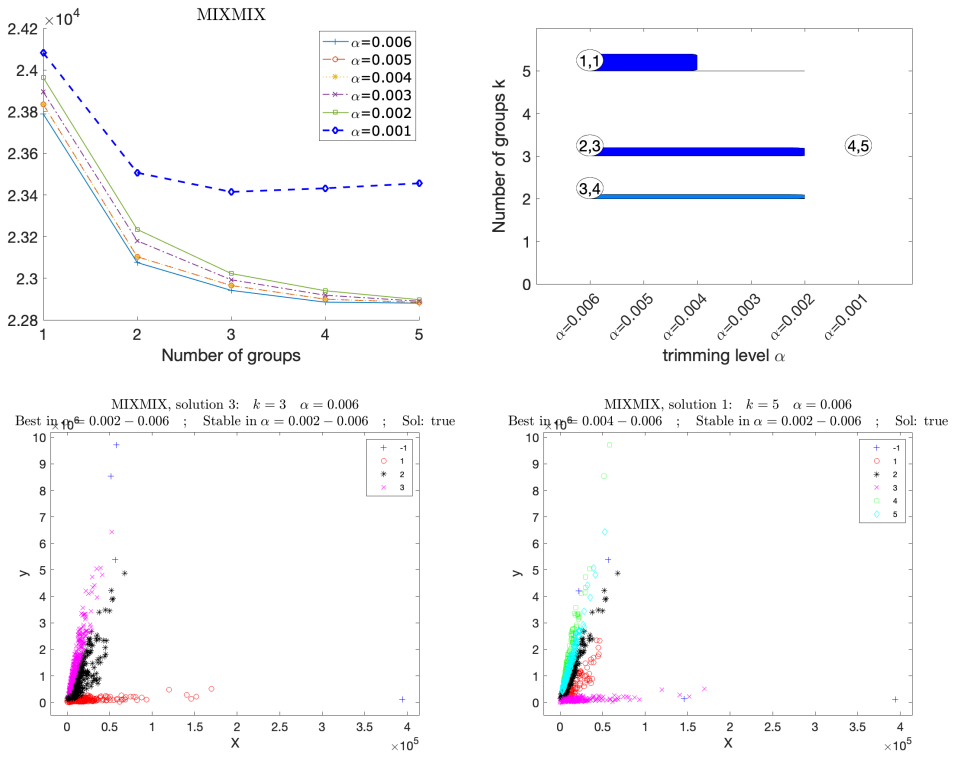


Figure 10: TCLUST-REG on the thinned sample; monitoring G and α for $c = 64$. Prices in €/ Kg estimated without intercept: for $G = 3$ we get $\{2.77, 35.16, 94.42\}$; for $G = 5$ we get $\{2.77, 26.21, 40.77, 93.11, 169.18\}$

.

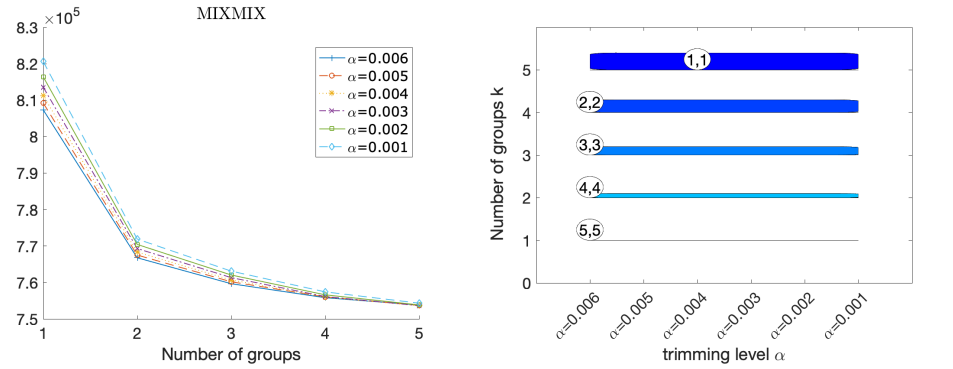


Figure 11: TCLUST-REG on the sample of 30,000 records; monitoring G and α for $c = 64$.

5 Detection of extreme outliers: the Surveillance monitoring system

The JRC is developing a *Surveillance Monitoring System* (SMS) meant to ensure the quality of the data and checking how they flow over time from the national systems into the DG TAXUD database. Section 6 comments on the frequency and updating of the data collection process, which is important to correctly interpret the information conveyed by any analysis based on this data source.

The system is based on an Oracle database and SqL procedures integrated with R functions. Results are published on a Tableau server. Its main purpose is to spot sudden changes in the reported data, which could appear as strong outliers or become structural. The outlying data records are flagged and this information is used by the other reporting systems of the JRC to ensure sound conclusions. In fact, while for complex estimations like those in the previous sections the JRC uses robust methods that resist to the presence of these anomalies, in a reporting or visualisation system based on the raw data, the aggregates or derived summary statistics could be easily biased.

Figure 12 shows an SMS dashboard that monitors the weekly aggregates of the import/export values declared at the EU Customs. Similar dashboards monitor the quantities (in kg) traded and supplementary units when available. The figure refers to a commodity that is close to those in our use case, which could be used to trade face masks. The increase during the COVID emergency (March-May 2020, in the green-

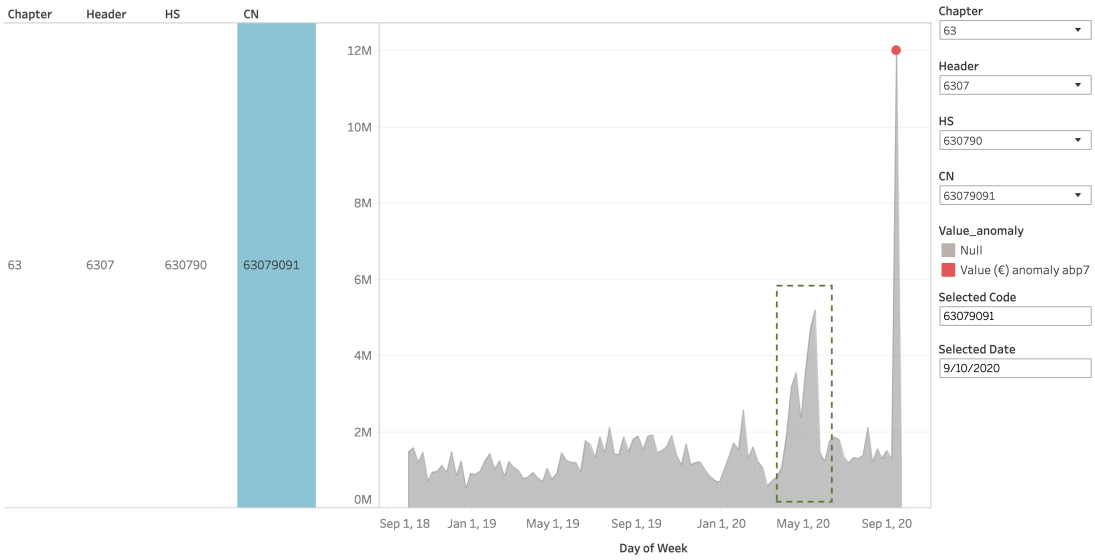


Figure 12: Surveillance database, CN code 6307.90.91: monitoring the weekly aggregates of the imported values in the period September 2018 -- September 2020. There is a structural increase due to the COVID emergency in the green-dotted rectangle and a visible outlier in the last month (red-filled bullet).

dotted rectangle) is in fact quite visible, but what is really striking is the red-filled bullet in the week of 13 September 2020. The Commission's trade policy is engaged to ensure an appropriate number of face masks in the EU during the emergency and more generally to build a solid policy for its market autonomy; it is therefore obvious that to determine whether such type of jumps are genuine or not is crucial.

To understand if the red point is the result of a clerical error or a more complex event (e.g. a series of large genuine imports), we can visualise a scatterplot of the quantities and values of the records reported in that particular week. Figure 13 plots three views of this type. The first on the top right refers to all records in the week. By clicking on the visible outlier (which accounts for more than €10 millions and 100,000 kg) it is possible to explode in a separate plot the data reported in the particular day which generated the outlier (15 September), as shown in the scatter on the bottom left. There is a single record that appears to have generated the weekly outlier: a click on it produces the information in the box of the same figure extracted from the Surveillance database. The plot at the bottom-right shows how the weekly data appear if the outlier is removed: they range now in reasonable and rather homogeneous value/quantity intervals. Outliers of this type are detected and flagged automatically following this on-line monitoring workflow:

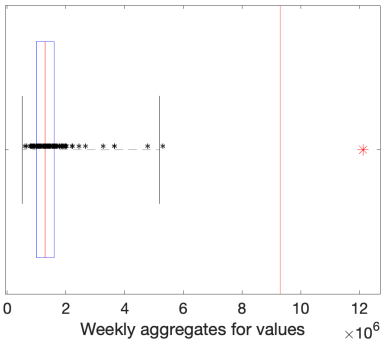
Collection. Every day, over night, the JRC downloads new data (including rectifications) from the official Surveillance database of DG TAXUD, and stores them in the local Oracle database.

Aggregation. Daily and weekly aggregates are built and stored in a separate table of the database.

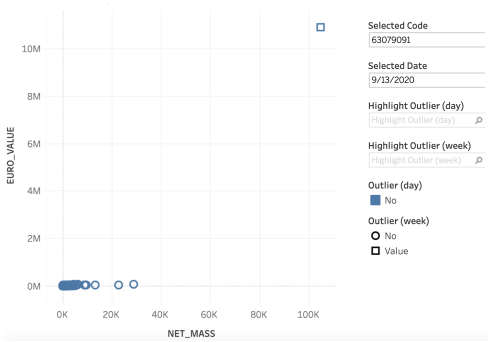
Weekly aggregates analysis. At the beginning of each week (Sunday), outliers are detected in time series of weekly aggregates of values and quantities (weights/units). The approach used for this purpose is based on an univariate boxplot adjusted for skewness (Hubert and Vandervieren, 2008) illustrated in Section C. The sliding time window covers the previous 2 years, therefore the points considered are 108 if all weeks are covered. The top-left panel of Figure 13 shows the boxplot of the weekly aggregates of Figure 12. Note the vertical black and red lines: the former refer to a classic boxplot, while the latter to the version adjusted for skewness.

Weekly data extraction. The anomalous cases spotted in the weekly analysis are further explored. The data (values or weights or units) for a particular week are extracted. The time period is enlarged if there are less than 50 points available, to ensure sufficiently reliable results.

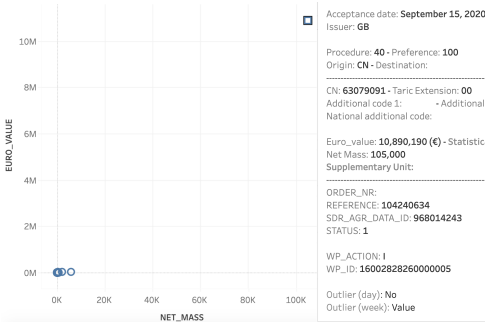
Weekly data sampling. The records reported in a particular week can be very high. In the case of commodity with CN code 63079098, which is closer to the face masks market, this number in the 2 years considered varies between 6,000 (week of 22/12/2019) and 68,000 (week of 12/04/2020). For other commodities there could be hundreds of thousands of records to treat and this requires the application of two subsequent sampling steps.



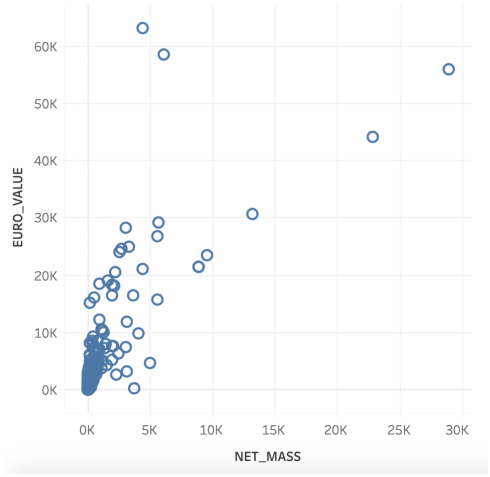
Boxplot adjusted for skewness for the 108 weekly aggregates of Figure 12, with the extreme outlier highlighted with a red asterisk.



All transactions with acceptance date in the same outlier's week (Sunday to Saturday, of 09/13/2020). The info box on the right refers to the entire scatter.



All transactions with acceptance date 09/13/2020. The info box on the right refers to the outlier, represented in the scatter with a blue square.



All transactions with acceptance date in the same week of 09/13/2020, without the outlier.

Figure 13: Surveillance database. CN code 6307.90.91. Upper left panel: boxplot representation of Figure 12. Upper right and bottom left panels: scatter plots of weekly and daily aggregates in the week of 9 September 2020 and on 13 September 2020. Lower right panel: weekly view excluding the outlier that is well visible as blue square symbol. These three scatter plots explain the origin of the peak in the time series of Figure 12 and in the corresponding boxplot.

1. If the dataset contains more that m (say $m = 1000$) records, then m records are selected and retained using weighted random sampling. This is done using FSDA function `randsampleFS`, documented at `http://rosa.unipr.it/FSDA/randsampleFS.html`. The vector of weights is calculated as the Euclidean norm of the `(weight,value)` pairs standardised by their respective maximum. Weighting sampling through the norm discharges only records that contain meaningless values and weights, which are extremely small.
2. The reduced sample obtained after the first sampling step can contain a large number of concentrated `(weight,value)` pairs, generated by trade operations very similar in terms of declared values and weights. We therefore apply the methodology discussed in Section B.3, to retain all meaningful records and only a small representative sub-sample of the `(weight,value)` pairs possibly present in the concentrated trade areas. The final number of sampled records is not known in advance.

Weekly data analysis. Univariate outliers are detected in the reduced sample of value declared in the selected week, again with a boxplot adjusted for skewness.

This machinery could be further developed in an operational system that uses the upper fence of the boxplot for classifying provisionally as outlier or genuine any new record coming during the next monitored week.

6 About declarations inflows and their timeliness

DG TAXUD receives the data on a daily basis, but Member States need some time to transfer the data to the Commission. The delay varies from country to country. In addition, customs authorities can rectify wrong data at any time, therefore the views in our monitoring dashboards are necessarily dynamic. For this reason, the procedures to compute all relevant statistics and to identify the outliers are run regularly: at the time of this writing once per week. For the moment we do not keep a record of the past views, therefore a record which appears outlying today could disappear in 1 week time or more.

Figure 14 shows some statistics on the data collection process. It is clear from the top panel that 40% of the Surveillance records are registered within the first week. Approximately another 40% enters the DG TAXUD database within 5 weeks. A final 5% has a delay bigger than 6 weeks. Note that the delay is computed from the day of last update (to account for all record corrections) and the acceptance date. In conclusion, we can say that the Surveillance dataset stabilises almost completely within a couple of months. This is in line with the Eurostat practice of publishing the official international EU trade statistics in their COMEXT database with a delay of three months.

There are other reasons for monitoring how data flow into the Surveillance database. The top panel of Figure 15 shows the progression of the monthly number of declarations of a specific commodity transmitted by a certain Member State to DG TAXUD. The number slowly increases until April 2020, passing from 1 to 10 millions records per month, then suddenly there is a drastic drop down to about 10,000 records per month, as it appears in the the zoom of the central panel. On the other hand, the bottom panel shows a rather stable progression of the corresponding value declared for the full set of monthly declarations, between €100 and 150 million per month. This change of pattern, which is confounding the nature of the trade dynamics, could be possibly explained by the possibility that one or several imports of many identical commodities have been actually split into a large number of declarations of much smaller quantity.

These two examples demonstrate the importance of implementing basic monitoring criteria.

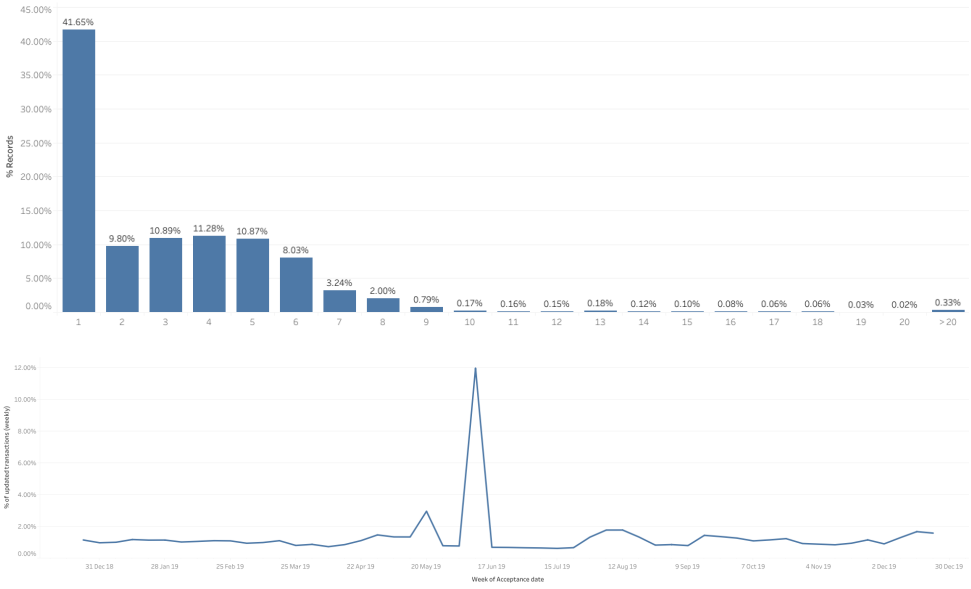


Figure 14: Monitoring the collection process of Surveillance data. Reference period is 01/01/2019 -- 31/12/2019. Top panel: Delay, measured in weeks, between the date of the registration of the customs declaration (acceptance date) and the date of the last update of the record. Bottom panel: Percentage of customs declarations that are updated on a weekly basis.

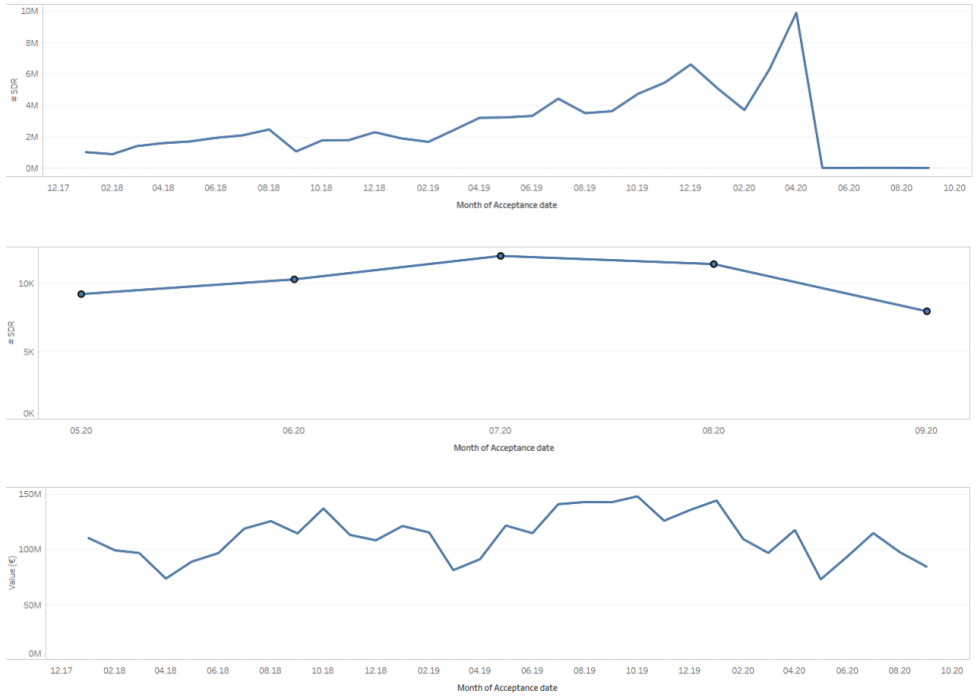


Figure 15: Monitoring the number of Single Data Records in Surveillance. Top panel: for this commodity, the number of records per month dropped from 10 million in April 2020 to 10.000 in May 2020. Middle panel: zoom of the period May - September 2020. Bottom panel: for the same commodity and the same period, the imported value per month remained in the range €70 million and €150 million. The Product code concerned has been masked for preserving confidentiality.

7 Conclusions

7.1 Policy outcome

In the COVID-19 crises JRC has established a Task Force in support to the Clearing House of the Secretariat General and other relevant services. We were requested to monitor the imports of various critical commodities in order to assess critical EU dependencies. In relation to face masks, we immediately realised that the task was complicated by the coexistence in a same TARIC code of various product types (FFP, surgical, etc) and other unrelated textiles. We were therefore forced to provide to the Clearing House statistics based on data mixing all face masks, regardless their type, but at the same time we could provide statistical evidence and information on the fine structure of these imports, with a an estimated price and weight breakdown. To better address this setback, JRC supported the proposal of DG TAXUD and ESTAT to open new TARIC codes for protective face masks. The proposal was adopted after long discussions among the relevant actors involved in finding a proper wording to describe the commodities of concern.

Almost contextually to the publication of this report, on 4th October 2020, the Customs Code Committee has introduced the new detailed codes for face masks, which we report in Table 5. On the JRC side, key success factors in this result have been:

— *A new methodological framework for customs data.* We have consolidated different statistical methods for the analysis of Surveillance data in a coherent analytic framework.

— *A new web-based tools for monitoring customs data.* We have implemented two web instruments for monitoring the European Union trade. One consists in a visualisation tool for presenting in suitable forms aggregated trade information (quantities and values) and related summary statistics. The second contains lists of signals (spikes, level shifts, trends) prioritised according to their statistical significance, associated with plots and charts that help evaluating in which trade context the signal has occurred.

7.2 Information gained on the trade of face masks

The visualization tool has allowed the drawing of conclusions of operational value during the COVID-19 crisis. For example, comparison of the trade flows in the year which preceded the pandemic (January 2019 – February 2020) and the first phase of the crisis (March 2020 – August 2020) revealed a dramatic increase

6307.90.98	—- Other:	
	—— Nonwovens (former 6307.90.98.10)	
	——— Protective face masks	
6307.90.98.11	———- FFP2 and FFP3 masks	p/st
6307.90.98.13	———- FFP1 masks	p/st
6307.90.98.15	———- Medical (surgical) face masks	p/st
6307.90.98.17	———- Other (Protective masks)	p/st
6307.90.98.19	———- Other (Nonwovens)	
	—— Other:	
6307.90.98.91	——— Hand-made	
6307.90.98.99	——— Other	

Table 5: New TARIC codes for face masks, valid as of 4th October 2020 (descriptions have been shorten: the correct ones can be consulted in TARIC). The table reflects the "Commission Implementing Regulation (EU) 2020/1369 of 29 September 2020, amending Annex I to Council Regulation (EEC) No 2658/87 on the tariff and statistical nomenclature and on the Common Customs Tariff".

of imported quantities and values for some of the analyzed commodities. We could easily spot that the weekly import of face masks (code 6307.90.98.10) jumped from an average of €3 Million and 580 tonnes per week to €410 million and 4180 tonnes per week, with an increase of the price from an average of 5 €/Kg to 100 €/Kg at the peak of the crisis. Other products have shown similar patterns (e.g. protective apparel for medical use, code 6210.10.92.00, or artificial ventilators, code 9019.20.00.00). For some of the commodities, we noticed a relevant change in the exporting countries market share toward an increased dependency of the EU from the import from China. For example, the share of EU import of face masks from China (value in €) jumped from 65% in the pre-COVID-19 phase to 97% in the COVID-19 period. This information turned out to be precious for the decisions of the Clearing House, in order to ensure the autonomy of the European Union for a wide set of commodities linked to the pandemic emergency.

7.3 Validation of the proposed methodology

The availability of the new codes in Table 5 will enable the validation of the approach for the identification of potential subgroups in Surveillance data proposed in this report. A natural way to proceed is to collect a number of declarations made after the adoption of the October 4th proposal, pool them together in a single dataset, and compare what our approach would predict on this dataset with what has been actually declared by the importers with the new codes. Note that the new coding introduces the possibility to declare the quantity imported in supplementary units. This will also allow complementary analyses focusing on the price per unit and the weight per unit estimation.

7.4 Deployment plan and next challenges

The JRC is now engaged in four main parallel activities.

— The refinement of the monitoring system. It is of interest to find what is the Member States composition in the samples extracted by weighted random sampling and the subsequent thinning step: we expect different patterns, for example Greek declarations under a million €, large number of declarations in Netherlands or declarations of large volume in Germany. The interpretation of the patterns and definition of the monitoring perspectives should be defined in collaboration with DG TAXUD.

— The consolidation of the platform that hosts our monitoring system, to ensure the service at any critical moment but also in view of its deployment in the operational infrastructures of DG TAXUD and OLAF. The key component of the IT platform is a very scalable Oracle database hosting the Surveillance data and other relevant datasets, which needs to respond in almost real time to address future operational monitoring needs. Currently our relational database is the largest one operating at the JRC.

— The development of appropriate statistical tests to decide automatically on the best model options.

For the TCLUST clustering, this means being able to decide if a dataset contains a single homogeneous group of data or multiple ones, how many the groups are, and if the variability of each group remains constant or not.

For the robust time series model, initially developed with Prof. Mia Hubert, Peter Rousseeuw (both at KU Leuven) and Marco Riani, the model selection consists of selecting the appropriate number and type of harmonics, level shifts and trend components. In this case the work is ongoing with the University of Pama (Marco Riani and Gianluca Morelli).

In both cases, the solution envisaged is expected to scan tens of thousands of datasets per week, therefore the computational efficiency of the solution is a crucial factor.

— The integration of different statistical modules into a unique integrated modelling framework. This is a methodological issue that requires the development of a combined testing approach.

References

Arnés-Novau, X., Checchi, E., Cerasa, A., Torti, F., Versino, C., Sevini, F. and Perrotta, D., 'Personal protective equipment (ppe) and other covid-19 related items: technical and trade control analysis', Tech. rep., European Commission, Joint Research Centre, 2020. Submitted.

Arsenis, S., Perrotta, D. and Torti, F., 'The estimation of fair prices of traded goods from outlier-free trade data', Tech. Rep. EUR 27696 EN, JRC-100018, European Commission, Joint Research Centre, Publications Office of the European Union, Luxembourg, 2015. ISBN 978-92-79-54576-4, doi:10.2788/3790.

Bowman, A. W. and Azzalini, A., 'Applied smoothing techniques for data analysis', Oxford University Press Inc, New York, 1997.

Brys, G., Hubert, M. and Struyf, A., 'A robust measure of skewness', *Journal of Computational and Graphical Statistics*, Vol. 13, No 4, 2004, pp. 996–1017.

Celeux, G. and Govaert, G., 'Gaussian parsimonious clustering models', *Pattern Recognition*, Vol. 28, No 5, 1995, pp. 781 – 793.

Cerioli, A., García-Escudero, L. A., Mayo-Iscar, A. and Riani, M., 'Finding the number of normal groups in model-based clustering via constrained likelihoods', *Journal of Computational and Graphical Statistics*, Vol. 27, No 2, 2018, pp. 404–416.

Cerioli, A. and Perrotta, D., 'Robust clustering around regression lines with high density regions', *Advances in Data Analysis and Classification*, Vol. 8, 2014, pp. 5–26. ISSN 1862-5347. . URL `http://dx.doi.org/10.1007/s11634-013-0151-5`.

EEC, 'Consolidated text: Commission Regulation (EEC) No 2454/93 of 2 July 1993 laying down provisions for the implementation of Council Regulation (EEC) No 2913/92 establishing the Community Customs Code', Tech. rep., European Commission, 1993.

EEC, 'Commission Implementing Regulation (EU) 2015/2447 of 24 November 2015, laying down detailed rules for implementing certain provisions of Regulation (EU) No 952/2013 of the European Parliament and of the Council laying down the Union Customs Code', Tech. rep., European Commission, 2015.

Eisenhauer, J. G., 'Regression through the origin', *Teaching Statistics*, Vol. 25, No 3, 2003, pp. 76–80. . URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9639.00136`.

Fraley, C. and Raftery, A. E., 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American Statistical Association*, Vol. 97, 2002, pp. 611–631.

García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A., 'A review of robust clustering methods', *Advances in Data Analysis and Classification*, Vol. 4, 2010a, pp. 89–109.

García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A. and San Martín, R., 'Robust clusterwise linear regression through trimming', *Computational Statistics & Data Analysis*, Vol. 54, No 12, 2010b, pp. 3057–3069.

Gordaliza, A., 'Best approximations to random variables based on trimming procedures', *Journal of Approximation Theory*, Vol. 64, No 2, 1991a, pp. 162 – 180. ISSN 0021-9045. .

Gordaliza, A., 'On the breakdown point of multivariate location estimators based on trimming procedures', *Statistics & Probability Letters*, Vol. 11, No 5, 1991b, pp. 387 – 394. ISSN 0167-7152. .

Hoaglin, D. C., Mosteller, F. and Tukey, J. W., eds., 'Understanding robust and exploratory data analysis', Wiley, New York, 1983.

Hubert, M. and Vandervieren, E., 'An adjusted boxplot for skewed distributions', *Computational Statistics & Data Analysis*, Vol. 52, No 12, 2008, pp. 5186 – 5201. ISSN 0167-9473. .

Megiddo, N. and Tamir, A., 'On the complexity of locating linear facilities in the plane', *Operations Research Letters*, Vol. 1, No 5, 1982, pp. 194 – 197. ISSN 0167-6377. . URL `http://www.sciencedirect.com/science/article/pii/0167637782900396`.

Perrotta, D. and Torti, F., 'Detecting price outliers in European trade data with the forward search'. In 'Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization', , edited by F. Palumbo, C. Lauro, and M. GreenacreSpringer-Verlag, Berlin, Heidelberg, 2010.

Perrotta, D., Torti, F., Cerasa, A. and Riani, M., 'The robust estimation of monthly prices of goods traded by the european union', Tech. Rep. EUR 30188 EN, JRC120407, European Commission, Joint Research Centre, Publications Office of the European Union, Luxembourg, 2020. ISBN 978-92-76-18351-8, doi:10.2760/635844.

Riani, M., Atkinson, A., Corbellini, A. and Laurini, F., 'Information criteria for outlier detection avoiding arbitrary significance levels', *Submitted*, 2020.

Riani, M., Corbellini, A. and Atkinson, A. C., 'The use of prior information in very robust regression for fraud detection', *International Statistical Review*, Vol. 86, 2018, p. 205– 218.

Riani, M. and Zani, S., 'An iterative method for the detection of multivariate outliers', *Metron*, Vol. 55, 1997, pp. 101–117.

Rousseeuw, P., Perrotta, D., Riani, M. and Hubert, M., 'Robust monitoring of time series with application to fraud detection', Vol. 9, 2019, pp. 108–121. ISSN 2452-3062. .

Rousseeuw, P. J. and Leroy, A. M., 'Robust regression and outlier detection', Wiley, New York, 1987.

Rousseeuw, P. J., Ruts, I. and Tukey, J. W., 'The bagplot: A bivariate boxplot', *American Statistician*, Vol. 53, 1990, pp. 87–88.

Silverman, B. W., 'Density estimation for statistics and data analysis', Chapman & Hall/CRC, 1986.

Torti, F., Perrotta, D., Riani, M. and Cerioli, A., 'Assessing trimming methodologies for clustering linear regression data', *Advances in Data Analysis and Classification*, Vol. 13, 2018, pp. 227–257. ISSN 1862-5347. .

Torti, F., Riani, M. and Morelli, G., 'Semiautomatic robust regression clustering of international trade data', *Submitted*, 2020.

Tukey, J. W., 'Exploratory data analysis', Addison-Wesley, Reading, Mass., 1977.

List of abbreviations and definitions

AMT Automatic Monitoring Tools

COMEXT Eurostat reference database for international trade in goods

CN Combined Nomenclature

EC European Commission

EU European Union

FFP Filtering Face-Piece

KDE Kernel Density Estimation

HS Harmonised System

JRC Joint Research Centre

OLAF Office de Lutte Anti-Fraude, European Anti-Fraud Office

RCLR Robust Clusterwise Linear Regression

SMS Surveillance Monitoring System

TCLUST-REG Trimmed Clustering in Regression

TARIC Integrated Tariff of the European Union

SAD Single Administrative Document

List of figures

List of tables

TECHNICAL APPENDICES

A Kernel Density Estimation (KDE) of unit prices

We start from a set of n import declarations for a given commodity of interest. For a generic import declaration $i \in \{1, \ldots, n\}$ we use the declared quantity x_i and statistical values y_i to build the *unit price* $u_i = y_i/x_i$. Then, a price distribution for the commodity of interest is obtained through the application of a kernel density estimation to the natural logarithm of the unit prices

$$l_i = \log(u_i) = \log\left(\frac{y_i}{x_i}\right). \tag{1}$$

The log transformation is applied to reduce skewness before estimating the density. The kernel distribution as usual is defined by a smoothing function and a bandwidth value, which control the smoothness of the resulting density curve. For this, we use the MATLAB kernel smoothing function `ksdensity.m` with the default normal kernel and optimal bandwidth for normal densities (Silverman, 1986, rule of thumb is applied to estimate the bandwidth). The density is evaluated at 1000 equally spaced points, within the limits found by `ksdensity.m` using a boundary correction that augments the support of the observed data (extensive treatment of these aspects can be found in Bowman and Azzalini, 1997).

In order to give more importance to those unit prices that involve a high quantity and/or statistical value, each declaration is weighted proportionally to its 'market share' according to the following expression:

$$w_i = \sqrt{\tilde{y}_i^2 + \tilde{x}_i^2} \quad \text{where} \quad \tilde{y}_i = \frac{y_i}{\sum_j y_j} \text{ and } \tilde{x}_i = \frac{x_i}{\sum_j x_j} \tag{2}$$

The weights on unit prices are incorporated in the density calculation using `ksdensity.m` option `'Weights',w`. Formally, the kernel function for a bandwidth h and normal kernel N, evaluated at a generic log-price value l, is expressed as

$$\hat{f}_h(l) = \frac{1}{h}\sum_{i=1}^{n} w_i N\left(\frac{l - l_i}{h}\right)$$

while the standard un-weighted form would be obtained with $w_i = 1/n$.

The estimated kernel density can have multiple local maxima, which form our set of *estimated modal prices*. The peaks are easily identified with standard optimisation methods. However, some of them might be spurious or irrelevant from the practical point of view. In order to focus on the most significant ones, we introduced these additional subsequent criteria:

— First we doubled the optimal bandwidth of Silverman (1986): this empirical over-smoothing criterion has proved to prevent from detecting irrelevant modal prices.

— Then we considered significant only modal prices with an area below the bell identified by the adjacent minimum points covering at least 5% of the total kernel density area. In practice, this means that we disregard prices that account for less than 5% of the market.

B Robust cluster-wise linear regression (RCLR) of values and weights

This approach is based on a general cluster-wise linear regression framework where we have n bivariate observations (y_i, x_i) with y the response (dependent variable) and x the explanatory (independent) variable. As in Approach A, the y_i represent the (statistical) values and x_i the quantity declared at the customs and recorded in the Surveillance database. Therefore, $x_i, y_i \geq 0$. Given that the trade quantities x_i are given, that is are assumed without error, the regressor x is not a random variable: we are under a *fixed regressors model*.

B.1 Contamination model

The model for the meaningful customs declarations can be represented as a mixture of G components, one for each potential sub-product in the commodity of interest:

$$f_0(y_i, x_i) = \sum_{g=1}^{G} \pi_g h_{y|x}(y_i, x_i, \beta_{0g}, \beta_{1g}, \sigma_g^2) \qquad \text{where:} \tag{3}$$

β_{0g} and β_{1g} are regression parameters specific to each mixture component, that is, the declared values depend linearly from the declared quantity as $y_i = \beta_{0g} - \beta_{1g} x_i + \epsilon_{ig}$. The error terms ϵ_{ig} capture all factors that influence y_i other than the x_i (the "noise").

σ_g^2 is the error variance within group g, which we assume constant. This means that σ_g^2, the variance of the errors ϵ_{ig}, remains the same regardless of the values of the x_i: we say that the responses y_i are homoscedastic.

π_g is the probability to observe a trade declaration in the group/component g;

We also assume that the errors are normally distributed. Therefore, conditionally on the observed quantity x, each mixture component can be expressed with the well known relation

$$h_{y|x}(y_i, x_i, \beta_{0g}, \beta_{1g}, \sigma_g^2) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left\{ -\frac{(y_i - \beta_{0g} - \beta_{1g} x_i)^2}{2\sigma_g^2} \right\}. \tag{4}$$

Under this common model, we have to compute reliable estimates of the parameters β_{0g}, β_{1g} and σ_g^2 in (3) for a given number of groups G. Then, we use the parameter estimates to assign each declaration to one of the G mixture components. For example, a unit can be assigned to the regression line that minimises the estimate of the scaled residual:

$$e_i = \frac{(y_i - \beta_{0g} - \beta_{1g} x_i)}{\sigma_g} \quad i = 1, 2, \ldots, n$$

As explained in Section 4, we may apply this approach also with the assumption that the linear approximation of the trade data generating process is forced to go through the origin, that is we may ignore the intercept term β_{0g} in the model.

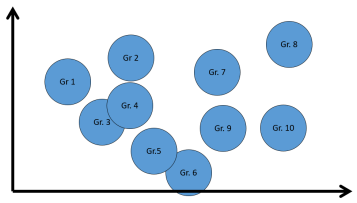
Given that customs declarations contain a lot of anomalous entries, an unknown number of pairs (y_i, x_i) cannot come from $f_0(y_i, x_i)$ in (3) but from an alternative distribution $c(y_i, x_i)$, the so called contaminant distribution. Thus, our model for the data could be better written as:

$$f_1(y_i, x_i) = (1 - \tau_c) f_0(y_i, x_i) + \tau_c c(y_i, x_i), \tag{5}$$

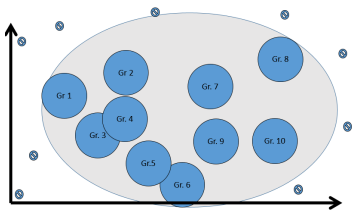
where $\tau_c < 0.5$ is the unknown contamination rate concerning observations of two types:

1. *Extreme data anomalies* that can exert a strong influence in the estimation process.
2. Departures from the underlying mixture model, giving rise to data points positioned between two groups; these *intermediate outliers* can influence the identification of the mixture components in (3) as much as the extreme ones.

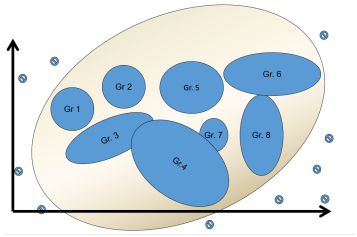
In practice, in most of the cases we expect relatively small values of τ_c, say $0 < \tau_c < 0.05$, because atypical but systematic trade operations can often be captured by adding a few more linear components in (3).



k-means: spherical groups of similar sizes, no outliers.



Trimmed k-means: spherical groups of similar sizes and outliers.



TCLUST: elliptical groups of different sizes and outliers.

Figure 16: Informal illustration of three multivariate model-based clustering approaches. TCLUST can control the cluster geometry, volume and eccentricity by constraining the covariance matrix. On the contrary, the traditional and widely used k-means implicitly assumes identical spherical groups (left panel). The spherical groups are because each variable (two in this case) has the same variance. If the variance in each dimension is allowed to vary (with a co-variance matrix constrained to be diagonal) we get an elliptical distribution (see Celeux and Govaert, 1995).

B.2 TCUST-REG

We need robust techniques to obtain reliable estimates of the parameter values under the contamination model (5), as the classical methods for clusterwise linear regression are usually sensitive to the outlying data generated by $c(y_i, x_i)$. An obvious way to address the problem is to avoid fitting the anomalous data, by leaving a proportion α of the most outlying observations unclassified: we say that these observations are *trimmed*. It is now well established that observations should be trimmed on the basis of the joint structure of the data, as discussed by Rousseeuw and Leroy (1987) and Gordaliza (1991a,b). The state-of-the-art implementation of the approach in model-based clustering is TCLUST (García-Escudero et al., 2010a) and here we consider its regression version, TCLUST-REG (García-Escudero et al., 2010b). In this model each mixture component follows equation (4) and the parameters that maximise the likelihood

$$\prod_{g=1}^{G} \prod_{i \in R_g} \pi_g h_{y|x(g)}(y_i | x_i, \beta_{0g}, \beta_{1g}, \sigma_g^2) \tag{6}$$

classify uniquely each observation into G non-overlapping groups $R_1, \ldots, R_G$, where R_g contains the indexes of the observations which are assigned to group g, and π_g is an unknown weight taking into account the group size([4]). The parameter values that maximise the likelihood are found with an Expectation Maximisation method with two constraints:

1. The first is to disregard the αn observations with the lowest contribution to the likelihood, that is $|\cup_{g=1}^{G} R_g| = n(1-\alpha) = h < n$. In practice, the n units are initially allocated to a cluster according to

$$f_{\max}(y_i, x_i) = \max_{g=1,\ldots,G} \pi_g h_{y|x(g)}(y_i, x_i, \beta_{0g}, \beta_{1g}, \sigma_g^2) \quad i = 1 \ldots n;$$

   then, these n numbers are ordered and the units associated with the smallest αn numbers are trimmed.

2. The second is to bound the likelihood and to avoid spurious groups, which is obtained by controlling the relative within-cluster variability with

$$\frac{\max_g \sigma_g}{\min_g \sigma_g} \leq c \tag{7}$$

   where $c \geq 1$ is a predefined maximum eccentricity constant called *restriction factor*. The importance of this constraint is illustrated in Figure 16 with the more intuitive, and perhaps better known, multivariate counterpart.

This *fixed partition model* is also referred to as *crisp clustering*. Alternatively we can consider a likelihood function that treats the cluster membership of each observation as random, as it should be for a *mixture model*, and assigns each observation to the cluster to which it is most likely to belong according to

$$\prod_{i=1}^{n} \left[\sum_{g=1}^{G} \pi_g h_{y|x(g)}(y_i | x_i, \beta_{0g}, \beta_{1g}, \sigma_g^2) \right] \tag{8}$$

([4]) The classification task in itself is a combinatorial partitioning problem that is NP-hard (see for example Megiddo and Tamir, 1982). TCLUST can be seen as a computationally treatable approximation to the problem.

where $f_0(y_i, x_i) = \sum_{g=1}^{G} \pi_g h_{y|x(g)}(y_i|x_i, \beta_{0g}, \beta_{1g}, \sigma_g^2)$ is the mixture density (3) from which the data are assumed to come and now π_g is a probability that an observation belongs to the mixture component g; being π_g probabilities, here we must have $\pi_g \geq 0$; $\sum_{g=1}^{G} \pi_g = 1$. In practice, the $h = n(1-\alpha)$ units to keep and classify are those which give the largest contribution to the likelihood (8), that is the h largest values of $f_0(y_i, x_i)$.

Our implementation of TCLUST-REG in the FSDA toolbox, function `tclustreg`, allows choosing between crisp assignment and mixture modelling through option `mixt`, set respectively to 1 and 2. Detailed documentation is at `http://rosa.unipr.it/FSDA/tclustreg.html`.

B.3 Dealing with concentrated samples in very large datasets: the 'small trade area'

Our experience is that even robust techniques like TCLUST-REG can fail when a large part of observations fall in a small region of the data space. This is the case with Surveillance data, as it occurs frequently that a large amount of declarations are small both in quantity and value: we call this high-density region of the data space 'the small trade area'. If this occurs, the effect of a high-density region on the estimation method is so strong that it can override the benefits of trimming and completely distort the estimates. In our case, this would mean ending up in totally unreliable price estimates. It should be noted that this problem potentially affects any other robust tool. For example, we had to address it for the detection of very extreme outliers (Section 5) with the boxplot adjusted for skewness briefly introduced in Section C.

We address the problem with an approach that we introduced in Cerioli and Perrotta (2014), consisting of sampling a small subset of observations which preserves the cluster structure of (3). The robust fitting methods are then applied only to the retained data. Note that the sampling is such that it retains also the main outliers that are generated by the contaminant $c(\cdot)$ in (5): in fact, while it is crucial that outliers do not influence the parameter estimates in (4), it is also important that the method is able to highlight them, because they may provide information about major anomalies like potentially fraudulent transactions. Although the discarded observations do not enter in the statistical assignment step, they could be set aside for further inspection at a subsequent stage.

This goal is achieved by defining a retention probability of each point as an inverse function of the estimated density function for the whole data set. In this extended framework, the contamination model (5) with the inclusion of a noise component for the 'small trade area' becomes:

$$f_2(y_i, x_i) = \tau_0 f_0(y_i, x_i) + \tau_c c(y_i, x_i) + \tau_d d(y_i, x_i), \tag{9}$$

with $\tau_0 > \tau_c$ (typically $\tau_0 \gg \tau_c$) and $\tau_0 + \tau_c + \tau_d = 1$. For $y_i, x_i > 0$, the noise component density $d(y_i, x_i)$ should be a smooth decreasing function of both $|y_i - \theta_y|$ and $|x_i - \theta_x|$. The center (θ_y, θ_x) corresponds to an unknown point in the scatter plot of (y_i, x_i), $i = 1, \ldots, n$, typically close to the origin. Furthermore, we expect that $\tau_d \gg \tau_c$.

In the FSDA toolbox this particular sampling step is implemented in the function `wthin`, with detailed documentation at `http://rosa.unipr.it/FSDA/wthin.html`. Function `tclustreg` applies automatically `wthin` with the option pair `<'wtrim',4>`.

B.4 Choosing the clustering hyperparameters through monitoring

Given the high number of different datasets to analyse (typically, one for each combination of product and origin), the number of components G must be estimated automatically. The same happens for the trimming level α and the restriction factor c in (7). We can say that these are the *hyperparameters* of the mixture model or associated clustering algorithm.

The suitability of a specific hyperparameter combination can be estimated with a statistic equivalent to the Bayesian Information Criterion (BIC) proposed for the mixture model by Fraley and Raftery (2002) and discussed in this clustering context by Cerioli et al. (2018). The `tclustreg.m` function of our FSDA toolbox returns this statistic in its standard output `out.MIXMIX`. A recent distinctive feature of this implementation is in the way the restrictions on the standard deviations associated to the model parameters are calculated to account for the trimmed observations. This calculation is based on a duality between trimmed observations and added variables, along the *means shift model* theory (Riani et al., 2020).

In choosing a suitable hyperparameters combination, it is important to consider its stability. For example, a solution based on three groups that is best for a wide range of trimming proportions is to be preferred to a two-groups solution that is optimal only for a specific trimming value.

Torti et al. (2020) have proposed a method to choose the best hyperparameters configuration by monitoring the solutions produced by different combinations. At this stage, the approach can be considered

semi-automatic because, although the best solutions are identified in an automatic way by comparing a set of different hyperparameter combinations, a proper inferential test to assess the significance of the solutions is not yet available.

In the FSDA toolbox this monitoring function is called `tclustregIC`, documented in `http://rosa.unipr.it/FSDA/tclustregIC.html`.

C The boxplot adjusted for skewness

This section illustrates the technique used to detect extreme outliers in a sample of n univariate data values x_i, $i = 1, \ldots, n$. We assume that the sample size n is reasonable. In the setting considered in this report this means $50 \leq n \leq 1000$.

A famous graphical data analysis tool for exploring a univariate sample is the boxplot. In the original formulation of Tukey (1977), an observation x_i is classified as a *potential* outlier if it falls outside the *fence* interval, that is if:

$$x_i < Q_1 - k \cdot IQR \quad \text{or} \quad x_i > Q_3 + k \cdot IQR \tag{10}$$

where $IQR = Q_3 - Q_1$ is a robust measure of scale based on the first and third quartiles (Q_1 and Q_3). Tukey proposed $k = 1.5$ as a reasonable constant for mild deviations and $k = 3$ for the far ones. These criteria are sound if the data distribute rather symmetrically, but if the tails are thick then also good observations risk exceeding the thresholds (Hoaglin et al., 1983, discuss this possibility pp. 59-65). The same happens if the data are skewed in some direction. Therefore, strictly speaking, the boxplot highlights 'potential' outliers, unless the data are normally distributed; in this case, the boxplot fences for $k = 1.5$ are comparable to the three-sigma rule (precisely, they corresponds to 2.689σ).

We use a boxplot adjusted for skewness in the form proposed by Hubert and Vandervieren (2008), which measures the skewness with the so called *medcouple* (MC, Brys et al., 2004), defined as:

$$MC = \underset{x_i \leq Q_2 \leq x_j}{\text{median}}\, h(x_i, x_j) \tag{11}$$

where Q_2 is the second quartile (that is the sample median) and $h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}$. Note that $-1 \leq MC \leq 1$, and that if data are skewed to the right than MC is positive, it is negative if the skewness is on the left, and it is 0 if data is symmetric. The LIBRA toolbox for MATLAB (`http://wis.kuleuven.be/stat/robust`) and the R package `robustbase` contain efficient functions for calculating MC.

Hubert and Vandervieren (2008) redefine the fences on the basis of the medcouple, replacing the thresholds (10) with:

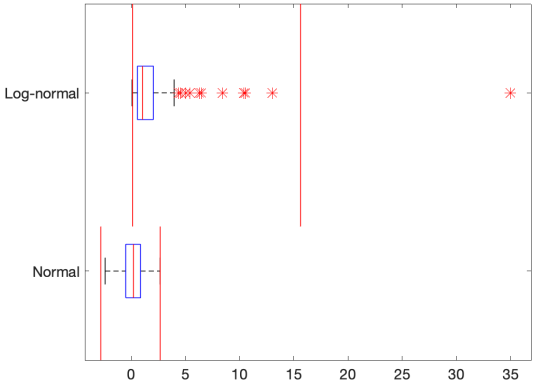
$$x_i < Q_1 - h_l(MC) \cdot IQR \quad \text{or} \quad x_i > Q_3 + h_u(MC) \cdot IQR \tag{12}$$

We found that with Surveillance data the best options for $h_l(MC)$ and $h_u MC)$, among those discussed by Hubert and Vandervieren (2008), are:

$$h_l(MC) = k \cdot \exp^{-4 \cdot MC} \quad \text{and} \quad h_u(MC) = k \cdot \exp^{3 \cdot MC} \tag{13}$$

Of course, given that Surveillance declarations are clearly skewed towards the larger values or weights and we are not interested in applying a lower threshold, we use only the upper fence.

The figure on the right exemplifies the use of the adjusted boxplot on two synthetic samples. There is a standard boxplot with superimposed the fences generated from a boxplot adjusted for skewness (red vertical lines). 100 data points are generated from a normal and log-normal distributions of 0-mean and standard deviation 1. The log-normal sample is contaminated with a single point of value 35. With normal data the adjusted fences practically coincide with the standard boxplot whiskers. In the case of the log-normal data, only the single contaminant is classified as outlier by the adjusted fences, while the traditional boxplot declares many more outliers.



As a final remark, it is possible to address the problem in a similar way also in the bivariate case, where we have n observations (y_i, x_i) with y representing the declared value and x the declared weight. In this case the tools that extend the boxplot are the *bagplot* by Rousseeuw et al. (1990), available in LIBRA, and the *bivariate ellipses* of Riani and Zani (1997), available in the FSDA toolbox in function `unibiv`, `http://rosa.unipr.it/FSDA/unibiv.html` or the bivariate boxplot in function `boxplotb`.

D Code to replicate the results of Section 4

```
%% monitor TCLUST-REG

rng(1234);            % for replication of results

load facemasks;       % load data from FSDA

values     = facemasks.data(:,2);          % dependent   variable
weights    = facemasks.data(:,1);          % independent variable
periods    = cell2mat(facemasks.rownames); % 0/1 pre-covid/covid

gscatter(weights,values,periods); % plot the data

% set variables for tclustreg
i_precovid = find(periods==0);
i_covid    = find(periods==1);
X = weights(i_covid);
y = values(i_covid);

% set parameters for tclustreg
typeIC    = 'MIXMIX';           % BIC criterion
alphaIni  = 0.006:-0.001:0.001; % monitored trimming percentages
alphaX    = 0;                  % no secon-level trimming
cc        = [1 4 16 32 64];     % possible restriction factors
intercept = false;              % model without intercept

% monitor tclust-reg
outIC = tclustregIC(y,X,'intercept',intercept, ...
                        'cc',cc(5),'whichIC',typeIC,...,
                        'alphaLik',alphaIni,'alphaX',alphaX,...
                        'nsamp',500,'plots',0);

% plot BIC information criterion as a function of c and k
tclustICplot(outIC,'whichIC','MIXMIX');

% extracts and plot a set of best relevant solutions
outICsol2 = tclustICsol(outIC,'whichIC',typeIC,'SpuriousSolutions',false);

% produces the carbike plot to find best relevant clustering solutions
carbikeplot(outICsol2,'SpuriousSolutions',false);

% run tclust-reg on the best solution
restrfact = cc(5); k=3; alphaLik=0.006; alphaX=0;
[out_3_6] = tclustreg(y,X,k,restrfact,alphaLik,alphaX,...
           'intercept',intercept,'nsamp',2000,'plots',1,'wtrim',0);

% run tclust-reg on the best solution
restrfact = cc(5); k=5; alphaLik=0.006; alphaX=0;
[out_5_6] = tclustreg(y,X,k,restrfact,alphaLik,alphaX,...
           'intercept',intercept,'nsamp',2000,'plots',1,'wtrim',0);
```

```
%% extract random sample based on weighted sampling: example
V_std = values/max(values);      % standardize y
W_std = weights/max(weights);    % standardize x
n     = numel(V_std);            % initial number of observations
kk = 100;                        % number of largest vectors to extract
w  = vecnorm([V_std,W_std]',2);  % compute the norm of the vector
ii = randsampleFS(n,kk,w(:));    % weighted random sample
V_sample = V_std(ii);            % extracted sample
W_sample = W_std(ii);

%% apply thinning: example
[Wt,pretain] = wthin([weights,values]);
W_t = weights(Wt);   % the sample retained
V_t = values(Wt);    % the sample retained
```

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-30431-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/817681

ISBN 978-92-76-24707-4