European Commission

JRC TECHNICAL REPORT

The Robust Estimation of Monthly Prices of Goods Traded by the European Union

A technical guide

Perrotta D., Cerasa A., Torti F. and Riani M.

2020

Joint Research Centre

EUR 30188 EN

Contact information
Name: Domenico Perrotta
Address: Joint Research Centre, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
E-mail: domenico.perrotta@ec.europa.eu
Tel.: +39 0332 785140

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

JRC120407

EUR 30188 EN

PDF	ISBN 978-92-76-18351-8	ISSN 1831-9424	doi:10.2760/635844
Print	ISBN 978-92-76-18352-5	ISSN 1018-5593	doi:10.2760/00361

Luxembourg: Publications Office of the European Union, 2020

How to cite this report: Perrotta, D., Cerasa, A., Torti, F. and Riani, M., *The Robust Estimation of Monthly Prices of Goods Traded by the European Union*, EUR 30188 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-18351-8, doi:10.2760/635844, JRC120407.

Table of contents

List of figures

Abstract

The general problem addressed in this document is the *estimation of "fair" import prices* from international trade data. The work is in support to the determination of the customs value at the moment of the customs formalities, to establish how much duty the importer must pay, and the post-clearance checks of individual transactions. The proposed approach can be naturally extended to the analysis of export flows and used for other purposes, including general market analyses.

The Joint Research Centre of the European Commission has previously addressed (Arsenis *et al.*, 2015) the trade price estimation problem by considering data for fixed product, origin and destination over a multi-annual time period, typically of 3 or 4 years, leading to price estimates that are specific for each EU Member State.

This report illustrates a different model whereby each price estimate is calculated on a monthly basis, using data for fixed time (month), product and origin. The approach differentiates between trades originated from different third countries and it is therefore particularly useful to monitor trends and anomalies in specific EU trade markets. These *Estimated European Monthly Prices* are publishes every month by the Joint Research Centre in a dedicated section of the THESEUS website (`https://theseus.jrc.ec.europa.eu`), accessible by authorized users of the EU and Member States services. The section, called *Monthly Fair Prices*, also shows the time evolution of worldwide price estimates computed with the same approach by fixing only time and product.

Acknowledgments

This work was partially supported by Administrative Arrangements of the "Automated Monitoring Tool" project (OLAF-JRC SI2.601156 and SI2.710969), funded under the Hercule III Programme. The AMT project has a long history, started with the leadership and visionary initiative of Spyros Arsenis in the first 2004 Hercule Programme, when statistical methods were at the margin of the Customs community.

Giuseppe Sgarlata has programmed or overseen all features and developments of the THESEUS website together with Marzia Grasso. They have also programmed the periodic automated downloads of COMEXT data and other relevant trade datasets (Surveillance 2 and 3), in addition to the regular production of the estimated prices disseminated by THESEUS. Massimiliano Gusmini is re-designing THESEUS under user-centered principles: the benefit of his contribution will be fully appreciated in the next releases of the web resource. Patrizia Calcaterra is ensuring the functioning of the service and its scalability to the drastic increase of users of the last few years.

The Forward Search method, used for the estimation of the "European Monthly Prices", was introduced by Anthony Atkinson (LSE), Marco Riani (University of Parma, co-author of this report) and Andrea Cerioli (University of Parma). The "Bayesian" extension of the Forward Search comes from joint work with Aldo Corbellini (University of Parma). Other statisticians are indirectly contributing to AMT, under the office of the Robust Statistics Academy of the University of Parma (`http://rosa.unipr.it/team.html`).

Thanks to Nicholas Shaw (OLAF) for commenting previous versions of the document. In OLAF several other persons contribute to AMT and THESEUS, but a special acknowledgment goes to Juergen Marke, who initiated many lines of work giving at the same time trust to the JRC explorations.

References to countries and products is made only for purposes of illustration and do not necessarily refer to cases investigated or under investigation by anti-fraud authorities.

1. Introduction

The rules of trade between nations are negotiated by the World Trade Organization (WTO)[1] in the spirit of *open trade*, "to ensure that trade flows as smoothly, predictably and freely as possible". The trade regulations are then enforced by the Customs system[2], the principal one being the collection of duties and taxes on imported goods. The majority of the Customs duties are established *ad valorem*, that is, the value of the imported goods is multiplied by the applicable duty rate to calculate the amount of duty payable at the moment of the Customs formalities. It is therefore essential for the Customs system to determine if the *value declared* by the importer is reasonably in line with the *value of the goods* traded.

Unfortunately the concept of "value" of a good has no unique definition and estimation method: economic theory has developed several competing schools of thought in this regard, where the concepts of *economic value*, *market value*, *market price* determined by a consumer and *trading price* are all different.

For this reason, the General Agreement on Tariffs and Trade (World Trade Organization, 1994) has established few general principles to determine the customs value of imported goods, introducing the concept of *actual value* of the imported merchandise[3]. Unfortunately the noble objective, aiming at a "fair, uniform and neutral system for the valuation that precludes the use of arbitrary or fictitious customs values", is hampered by the concrete applicability of the principles when there are reasons to doubt the accuracy of the declared value. In this regard, it has been noted (FATF, 2006)[4] that the difficulty of customs services "in identifying over and under invoicing and correctly assessing duties and taxes [is due] in part [...to the fact that] many customs agencies do not have access to data and resources to establish the *"fair" market price* of many goods".

This document proposes an approach that uses international trade data to estimate such "fair price" with monthly precision. The work goes in supports to the determination of the customs value, and is especially applicable during the post-clearance audits done by the Customs well after the clearance checks at the border[5]. The discussion is limited to the methodological aspects of the approach, not having the ambition to address the implementation of a price estimation approach under the customs value's legal framework. Our data-driven solution is based on solid statistical theory, and can be naturally extended to the analysis of export flows or used for other purposes, including general market analyses.

The rest of the report is organised as follows. The following section locates this work in a wider historical context, with a view on the term "fair price", its origin, current use and alternative formulations. Section 2 introduces the JRC approach to the estimated European monthly prices; to be coherent with what the Member States find in the anti-fraud resources of the EC (THESEUS and AFIS), the estimates and the model will be referred to as *European Monthly (Fair) Prices*. Sections 3 and 4 detail the general approach, the model and estimation method. Section 5 illustrates how the European Monthly (Fair) Prices are disseminated in THESEUS. The first Appendix, A, elaborates on the mathematical motivations for the choice and use of the term "Fair Prices", which is now familiar to the THESEUS community, but is source of controversy when it is linked to the WTO guiding principles of customs valuation. The other Appendices provide supplementary material on the methodological aspects of the estimation method.

2. Historical and terminological considerations

The European Commission (EC) introduced the term "fair price" in relation to Customs valuation during the common priority control area "Discount" exercise promoted in 2011 by the Directorate General for Taxation

[1] The WTO is a global economic policy-making organization that decides on the system of trade rules. Its member governments cover 98% of the world trade. WTO replaced in 1995 the General Agreement on Tariffs and Trade (GATT), established in 1948 to repair the disasters on international trade of the Second World War.

[2] The international Customs system is governed by the World Customs Organization (WCO), established in 1952 "to enhance the effectiveness and efficiency of Customs administrations". The national representatives in the WCO cover approximately 98% of world trade.

[3] The GATT defines the *actual value* as "the price at which, at a time and place determined by the legislation of the country of importation, such or like merchandise is sold or offered for sale in the ordinary course of trade under fully competitive conditions".

[4] The Financial Action Task Force (FATF) is an inter-governmental policy-making body established in 1989 "to set standards and promote effective implementation of legal, regulatory and operational measures for combating money laundering, terrorist financing and other related threats to the integrity of the international financial system". The FATF has developed a series of recommendations also to fight against Trade-Based Money Laundering (TBML), a practice used to move illicit funds through financial transactions associated with the trade in goods and services (FATF, 2006, 2008, 2012, 2013).

[5] Border controls at the point of clearance cannot be excessive and time-consuming and often rely on limited documentation, insufficient to properly determine the correct Customs value, good classification and real origin. Therefore, structured controls and examinations of individual transactions are often done after the importation/exportation at the frontier, where only selective and targeted checks can be done.

(TAXUD), and during the Joint Customs Operation "Snake" initiated by the EC Anti-Fraud Office (OLAF) in 2013. Both were targeting the undervaluation of textiles and footwear from Asian countries. The EC provided guidelines to Member States on how to tackle undervaluation by comparing declared prices with lists of "fair prices" disseminated after accurate selection by OLAF through its Anti-Fraud Information System (AFIS, AMT section[6]) and in comprehensive (unfiltered) form by the EC Joint Research Centre (JRC) through its THESEUS web resource (`https://theseus.jrc.ec.europa.eu`). While, for each commodity, THESEUS contained detailed country-specific estimates (one fair price for each Member State), AFIS focused on appropriate combinations of the THESEUS fair prices into a unique EU-wide *estimated European price*.

The methodology used to estimate the country-specific estimates based on data observed in a time window of 4-years, was detailed in a public JRC technical report by Arsenis *et al.* (2015). The report, primarily addressed to the THESEUS community, adopted the term "Fair Prices" for historical consistency, but also and especially for statistical/mathematical motivations, which are briefly illustrated in Annex A. These motivations have no relation with the fairness concept stated by the *guiding principles of customs valuation* in (World Trade Organization, 1994).

In 2016 the European Court of Auditors (ECA) carried out a performance audit to assess whether the Commission and the Member States have designed robust import procedures that protect the financial interests of the EU. In the report, the ECA made substantial reference to the fair price estimates in THESEUS but tried to go beyond the "fairness" concept, by re-defining them as *Outlier-Free Average Prices*[7], meaning that they are "statistical estimates calculated for the prices of traded products on the basis of outlier-free data". Technically speaking ECA's definition is correct, although the advocated "average" is in a form weighted by the trade quantities through a regression model (equation 3 of page 11 of Arsenis *et al.* (2015)) and the robust estimation methods used by the JRC do not require that the data are cleaned beforehand.

To summarize, the price estimates in Arsenis *et al.* (2015) and those discussed in this document should be read as *baseline values* for the import price (regardless the fact of being MS-specific or EU-wide). In operational contexts, these baseline values are associated with a *decision rule*, used as dividing line between regular trade and potential undervaluation.

3. Approach

The main data source considered in this report comprises monthly aggregates of trade values and quantities generated for the same product and partners in trade. These aggregates are downloaded from the COMEXT database of the European Statistical Office, Eurostat. In COMEXT the product codes are classified at the detailed 8-digits level of the Combined Nomenclature (CN8); therefore, for many commodities, the records referring to the same product code are reasonably homogeneous. The quantities are given in tons and supplementary units if foreseen, and the values in thousands of Euros. The method and model in the proposed approach are unaffected by the choice of the measurement unit used to represent quantities and values. Therefore, the approach can be naturally extended to more detailed daily aggregates or to single declaration level data, where quantities are typically expressed in Kilograms and values in Euro[8].

The JRC has already addressed the fair price estimation problem (Arsenis *et al.*, 2015) by considering COMEXT data for each Product (P), Origin (O) and Destination (D) taken over a multi-annual Time (T) period of 4 years. We indicate these datasets, formed by at most 48 data units, with POD*. Every month the JRC computes new POD* estimates based on the previous 4-years period, and disseminate them through the `Fair Prices section (FP)` of the THESEUS website, for the consultation of anti-fraud users in OLAF and the Member States. In THESEUS, the FP estimates are accompanied by scatter plots highlighting quantity-value flaws that are anomalous with respect to the FP, the so called "price outliers". These outliers may be due to under-valuation, but also recording errors or legitimate, though peculiar, market dynamics. The FP approach has two potential shortcomings:

- Time dependent effects such as seasonality or trend in the trade price over time are only partially taken into account.

[6] AMT, which stands for "Automated Monitoring Tool", is also a series of joint projects between OLAF and the JRC, supported by the HERCULE anti-fraud Programmes.

[7] In statistics, an **outlier** is a data point that differs significantly from other observations, due to measurement variability, recording errors, natural anomalies, but also (in relation to customs fraud) data manipulations or miss-declarations. Whatever the nature of the outlier is, its presence can cause serious distortions of statistical analysis (in our context, a severe distortion of the fair price). A reference when outliers are in relation to data following a linear regression structure, of interest for this report, is Rousseeuw and Leroy (1987).

[8] Customs collect the import/export declarations in the Single Administrative Declaration (SAD) form. The data are then transmitted to the EC in the Surveillance database of TAXUD. Other EC services can access the data, often with trader information removed for granting anonymity or with some form of aggregation (e.g. by day).

- In the European internal market there should not be distinction between the trade in different Member States.

There is therefore the need, for each combination of product and origin, of a reference EU price estimated over a reasonably short time period. This report addresses this need with a model whereby each price estimate is calculated on a monthly basis, using COMEXT aggregates for fixed Product, Origin and Time (PO*T). These *estimated European monthly prices* are disseminated by THESEUS in the section `Monthly Fair Prices` (MFP).

Obviously, a PO*T dataset can contain up to 28 data units (monthly COMEXT aggregates), one for each Member State in trade with a given Origin. Therefore, the practical applicability of the MFP model may be limited by the availability of a sufficient number of data units offered by COMEXT. The approach presented in this document addresses the potential small sample size problem with an "empirical Bayes" extension of a robust regression estimation method known as Forward Search (Atkinson and Riani, 2000). The Bayesian component comes from the fact that at a given month the estimation, if necessary, takes into account also previous estimates or data units from one or more preceding months. For this reason, the approach is potentially applicable even in extreme unfortunate cases, when for example only a single or no trade flow is available at a given month. In addition, the approach has the advantage of "smoothing" the estimated price series, in the sense of reducing the number of episodic fair price jumps, thus providing a representation that better captures the meaningful trade price trends.

4. Model, estimation and outlier detection

As in Arsenis *et al.* (2015), we assume that in a PO*T dataset, for data points that are not outliers, the monthly aggregated quantities (Q) are recorded without systematic errors, the monthly aggregated values (V) are recorded with errors and thus are related to what is called the linear regression with no intercept, that is:

$$V_{POT,d} = p_{POT} \cdot Q_{POT,d} + \epsilon_{POT,d} \tag{1}$$

where p_{POT} is the parameter to estimate and $\epsilon_{POT,d}$ are random, independent errors with zero mean and an unknown constant variance σ^2 for all observations in the dataset. As usual we do not exclude the presence of outliers in the data. An outlier is a data point of quantity and value (Q, V) that does not follow the distribution specified by the assumed regression line. The fair price is the slope of a regression line fit on a "clean", i.e. an outlier-free, set of data points.

For the sake of generality and to simplify the discussion, from now on we will use the usual notation for the independent and explanatory variables of a regression model, that is, y and X will take the place of $V_{POT,d}$ and $Q_{POT,d}$ respectively (the explanatory variable being in capital letter because in general it is a vector of different observed variables).

In the statistical literature there are two broad approaches to distinguish between the subset of "good" data points and the outliers (see, e.g., Barnett and Lewis, 1994; Rousseeuw and Leroy, 1987). The first approach uses *regression diagnostic* tools to suppress the outliers and fits the remaining (supposedly good) data by least squares; Arsenis *et al.* (2015) have discussed the FP model based on POD* data in THESEUS under an approach of this first family, with the introduction of appropriate corrections for multiple testing. The second approach fits a *robust regression* line on a fixed percentage of appropriately selected data (at least 50%) and then identifies as outliers the observations that deviate from the robust fit. The robust approach family includes many modern estimation methods, the most popular being the Least Median of Squares (LMS, Rousseeuw, 1984), Least Trimmed Squares (LTS, Rousseeuw and Van Driessen, 2006), M (Huber, 1973), S (Rousseeuw and Yohai, 1984), MM (Yohai, 1987) and the Forward Search (Atkinson and Riani, 2000). The JRC has implemented tools of both families in collaboration with its academic partners in the University of Parma, under the SAS and MATLAB platforms[9].

Among the robust methods, the *Forward Search (FS)* has the advantage of not fixing in advance the percentage of data units to be used for the model fit, because the method optimally adapts this number during the fitting process itself. Several experimental assessments (see, e.g. Torti *et al.*, 2012; Salini *et al.*, 2016) and studies (Riani *et al.*, 2014b) show that the FS approach produces superior performances: therefore, we use this approach as the basis for robust regression.

[9]The MATLAB implementation, called FSDA, consists of a very extensive set of robust tools that we distribute with full documentation at the website of the Interdepartmental Centre of Robust Statistics (Ro.S.A.) of the University of Parma `http://rosa.unipr.it/fsda.html`, in the MATLAB Central File Exchange `https://it.mathworks.com/matlabcentral/fileexchange/72999-fsda` and in GitHub `https://github.com/UniprJRC/FSDA`. Inside the European Commission the toolbox is presented at `http://fsda.jrc.ec.europa.eu`.

Let us see how the adaptive mechanism of the FS works. In order to detect outliers and departures from the fitted regression model, the FS in regression uses least squares to fit the model to subsets of m observations, starting from an initial subset of m_0 observations (in principle m_0, in our simple regression model, can be even equal to 1). The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest squared residuals. For each m ($m_0 \leq m \leq n-1$), we test for the presence of outliers, using the observation outside the subset with the smallest absolute deletion residual. We leave out of the discussion the details about the testing process, to not burden the reader with technical aspects that are extensively treated in the already cited Forward Search literature. Rather, we illustrate in Annex B with a simple example the advantages of the adaptive mechanism offered by the FS.

This works in the absence of any information on the past: only the observed data units are used in the process. However, our context offers appreciable prior information about the values of the parameters in a specific month (Perrotta and Torti, 2010). This can sometimes be thought of as coming from n_0 fictitious observations y_0 with matrix of explanatory variables X_0. Then the data consist of the n_0 fictitious observations plus n actual observations. This new procedure is called in the literature *Bayesian Forward Search* (Atkinson *et al.*, 2016a, 2017; Riani *et al.*, 2018), which can be seen as an "empirical Bayes"-type approximation of a canonical Bayesian approach, as the prior information is actually coming from past data.

The Bayesian FS process in this case now proceeds from $m = 0$, when the fictitious observations provide the parameter values for all n residuals from the data. The search then continues as outlined above but with the fictitious observations always included in those used for fitting, their residuals being ignored in the selection of successive subsets. The use of prior information based on previous months has enabled to treat also cases in which the number of observations is very small and even if when there is just a single observation for a particular month. The key components of the Bayesian FS are detailed in the Annexes from C to F.

We close the section with a final ingredient considered in our approach to the MFP estimation. Experience with COMEXT and especially Surveillance data has shown that it is important to complement the analysis with the indication of the R^2 index about the goodness of fit. A value too low for the R^2 is an indication about lack of fit of the model, due for example to the presence of multiple populations or to heteroskedasticity[10] (Atkinson *et al.*, 2016b). At the other extreme, a too high value of R^2 indicates the presence of a (almost) perfect fit and may lead to declare as outliers units which differ only slightly from the regression line constructed using the bulk of the data (the problem is mentioned in the book of Maronna *et al.* (2006) and is addressed in a practical context by Pagano *et al.* (2012)). In this case it is necessary to artificially modify (in our Bayesian context) the prior value of residual sum of squares. The details of this element of our approach are in Annex G.

5. Monthly Fair Prices in THESEUS

The use of the MFP section of THESEUS is detailed in a tutorial document (`https://theseus.jrc.ec.europa.eu/Report/TutorialMFP.pdf`). Here we briefly illustrate the main output produced by the MFP model with an example, based on data considered in the period January 2014 - October 2019 for a rather broad product category: "handbags with outer surface of plastics sheeting" (CN-code 42022210). Figure 1 shows the evolution of two monthly price estimates obtained as discussed in this document:

- The "benchmark price" (black line) is estimated considering, in each month, the trades of all origins and all destinations (the P**T estimate);
- The "origin price" (green line) focuses on a single origin (China, in this case) and all destinations (PO*T estimate, with O='CN').

The estimated price for CN is rather stable and lower than the world market price for most of the period. Figure 2 shows the corresponding table of estimated monthly prices in the time window considered. The estimated monthly price is complemented by a few key statistics, including robust confidence interval for the price estimates, goodness of fit statistic R^2, number of observations used for the fit, number of outliers possibly detected, and a quality index indicator measuring the reliability of the regression as a weighted average between the value of the goodness of fit and the number of observations. More precisely, the index is calculated according to the following formula:

$$QI = \frac{2R^2 + \#\text{Obs}/28}{3} \tag{2}$$

[10] *Heteroskedasticity* refers to sub-populations that have different variabilities from other data.

where 28 is the maximum number of monthly observations (the number of EU Member States). Since $0 \leq R^2 \leq 1$ and $0 \leq \#\text{Obs} \leq 28$, then $0 \leq QI \leq 1$. The index is graphically represented with up to 5 vertical colored bars. The bars in the figure mean that, in this case, the estimates are all very reliable.

Figure 3 shows scatter plots obtained by clicking on the graph icon on the right side of the table. The four panels refer respectively to data for November 2016, December 2016, January 2017 and February 2017. The plots show that an anatomized Member State (MS X) in the four consecutive months appears as low price outlier.

Finally, Figure 4 shows a table of the destination details on the right of the THESEUS section. Clicking on the lines for the MS X and FR, the flows of these two Member States are superimposed to the scatter plot of the observed vs estimated prices of Figure 1, giving rise to the plot of Figure 5, which now reveals (for a certain period of time) a systematic tendency to under-price in MS X and a perfect adherence with the estimated EU price (green line) for FR.

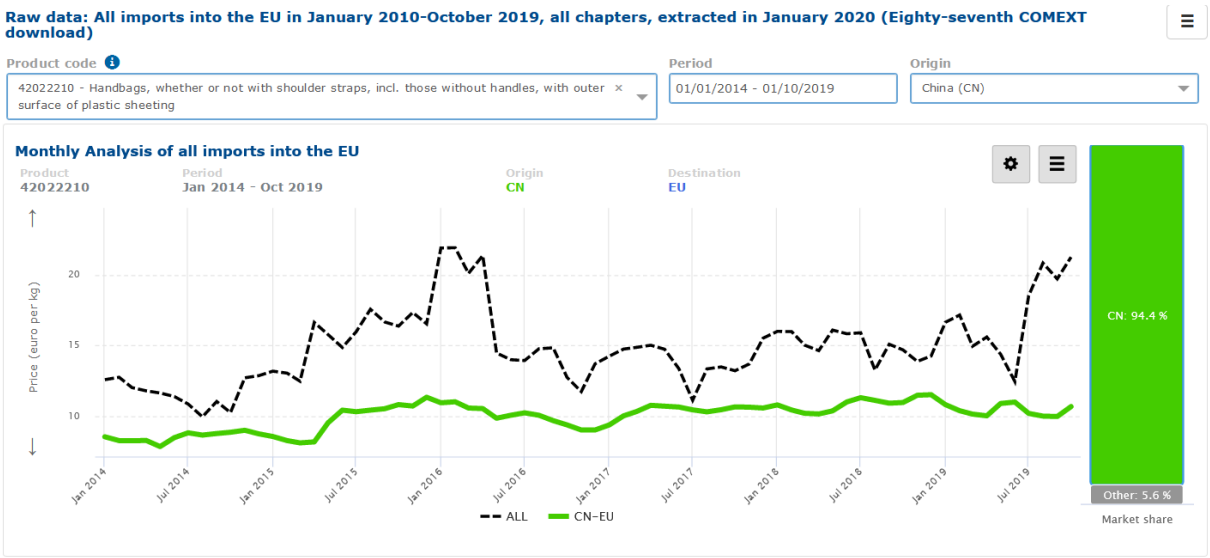

Figure 1: Historical evolution of the monthly price estimates in the period January 2014 - October 2019. Product code: 42022210, "handbags with outer surface of plastics sheeting". Black line: all origins; Green line: CN origin.

Period	Estimated fair price €/Kg	Estimated fair price interval €/Kg	Goodness of fit	Obs #	Quality index	Out #
November 2015	**10.72**	(10.12 ; 11.33)	0.961	26	▮▮▮▮▮	1
December 2015	**11.34**	(10.70 ; 11.99)	0.967	25	▮▮▮▮▮	2
January 2016	**10.95**	(10.32 ; 11.57)	0.953	25	▮▮▮▮▮	3
February 2016	**11.02**	(10.40 ; 11.64)	0.970	26	▮▮▮▮▮	2
March 2016	**10.58**	(9.98 ; 11.17)	0.961	27	▮▮▮▮▮	0
April 2016	**10.54**	(9.92 ; 11.16)	0.946	27	▮▮▮▮▮	1
May 2016	**9.86**	(9.29 ; 10.43)	0.953	24	▮▮▮▮▮	3

$0 \leq QI < 0.4$ →
$0.4 \leq QI < 0.7$ →
$0.7 \leq QI < 0.8$ →
$0.8 \leq QI < 0.9$ →
$0.9 \leq QI \leq 1$ →

Figure 2: Table of the monthly price estimates in the period November 2015 - May 2016. Product code: 42022210, "handbags with outer surface of plastics sheeting". Origin: China. On the right, the colour-map representing the quality index value.

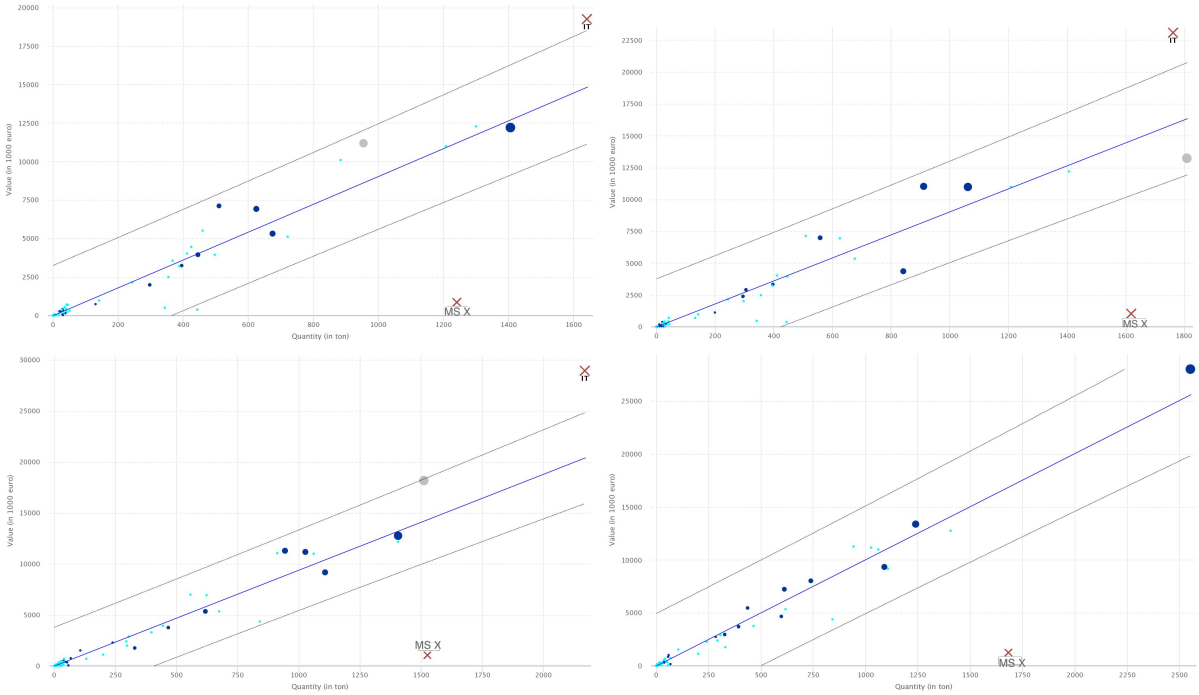

Figure 3: PO*T scatterplots for P='42022210', O='CN' and T = 'Nov 2016' (top-left), 'Dec 2016' (top-right), 'Jan 2017' (bottom-left) and 'Feb 2017' (bottom-right).

Observations by destination

Product	Period	Origin	Destination
42022210	Jan 2014 - Oct 2019	CN	EU

Destination	Obs #	Out+ #	Out- #	Market share %
Italy (IT)	70	21	0	18.1%
United Kingdom (GB)	70	0	8	14.3%
Spain (ES)	70	0	0	12.9%
France (FR)	70	0	0	7.6%
Germany (DE)	70	0	0	6.9%
MS X	70	0	29	6.2%
Belgium (BE)	70	0	0	6.1%

Figure 4: Table of observed vs estimated prices by Member State of destination in the period January 2014 - October 2019. Product code: 42022210, "handbags with outer surface of plastics sheeting". Origin: China. Only few Member States are visualized.

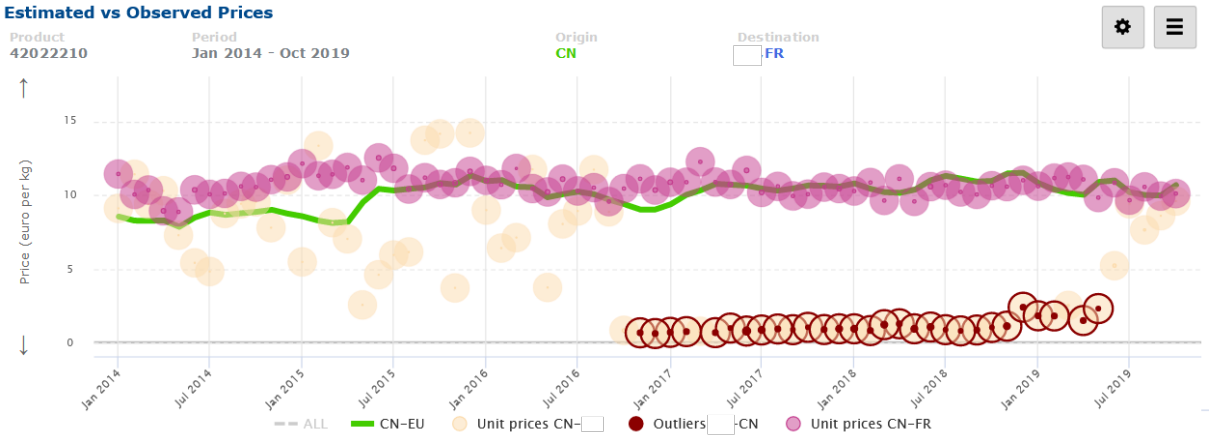

Figure 5: Plots of observed vs estimated prices, for an anonymous Member State X and France destinations. The green line is the estimated fair price for imports from China (PO*T estimate for P=42022210, O='CN' and T in the period January 2014 - October 2019). France is in perfect alignment with the estimated EU price. MS X in a certain period shows systematic under-pricing.

References

Arsenis, S., Perrotta, D., and Torti, F. (2015). The estimation of fair prices of traded goods from outlier-free trade data. Technical Report JRC-100018, European Commission, Joint Research Centre. EUR 27696 EN, ISBN 978-92-79-54576-4, doi:10.2788/3790.

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

Atkinson, A. C., Corbellini, A., and Riani, M. (2016a). Incorporating prior information into the forward search for regression. In A. Di Battista, E. Moreno, and W. Racugno, editors, *Topics on Methodological and Applied Statistical Inference, Studies in Theoretical and Applied Statistics*. Springer, Heidelburg. DOI 10.1007/978-3-319-44093-4_1.

Atkinson, A. C., Riani, M., and Torti, F. (2016b). Robust methods for heteroskedastic regression. *Computational Statistics and Data Analysis*, **104**, 209 – 222.

Atkinson, A. C., Corbellini, A., and Riani, M. (2017). Robust Bayesian regression with the forward search: Theory and data analysis. *TEST*, (26), 869?886.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data, 3rd edition*. Wiley, New York.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651–659.

FATF (2006). Trade based money laundering. Financial Action Task Force, www.fatf-gafi.org.

FATF (2008). Best practices paperbest practices on trade based money laundering. Financial Action Task Force, www.fatf-gafi.org.

FATF (2012). Apg typology report on trade based money laundering. Financial Action Task Force, www.fatf-gafi.org.

FATF (2013). Money laundering and terrorist financing through trade in diamonds. Financial Action Task Force, www.fatf-gafi.org.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, **1**, 799–821.

Johansen, S. and Nielsen, B. (2015). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli*, **21**. (In press).

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester.

Pagano, A., Perrotta, D., and Arsenis, S. (2012). Imputation and outlier detection in banking datasets. In *46th scientific meeting of the Italian Statistical Society*. CLEUP. ISBN 978-88-6129-882-8.

Perrotta, D. and Torti, F. (2010). Detecting price outliers in European trade data with the forward search. In F. Palumbo, C. Lauro, and M. Greenacre, editors, *Data Analysis and Classification*. Springer-Verlag, Heidelberg.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications, 2nd edition*. Wiley, New York.

Riani, M., Cerioli, A., Atkinson, A., Perrotta, D., and Torti, F. (2008). Fitting mixtures of regression lines with the forward search. In F. Fogelman-Soulié, D. Perrotta, J. Piskorski, and R. Steinberger, editors, *Mining Massive Data Sets for Security*, pages 271–286. IOS Press, Amsterdam.

Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.

Riani, M., Cerioli, A., and Torti, F. (2014a). On consistency factors and efficiency of robust S-estimators. *TEST*, **23**, 356–387.

Riani, M., Atkinson, A. C., and Perrotta, D. (2014b). A parametric framework for the comparison of methods of very robust regression. *Statistical Science*, **29**, 128–143.

Riani, M., Corbellini, A., and Atkinson, A. C. (2018). The use of prior information in very robust regression for fraud detection. *International Statistical Review*, **86**(2), 205–218.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Rousseeuw, P. J. and Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, **12**, 29–45.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis: Lecture Notes in Statistics 26*, pages 256–272. Springer Verlag, New York.

Salini, S., A., C., F., L., and M., R. (2016). Reliable robust regression diagnostics. *International Statistical review*, **84**(1), 99–127. doi:10.1111/insr.12103.

Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics*, **34**, 940–944.

Torti, F., Perrotta, D., Atkinson, A. C., and Riani, M. (2012). Benchmark testing of algorithms for very robust regression: FS, LMS and LTS. *Computational Statistics and Data Analysis*, **56**, 2501–2512. doi:10.1016/j.csda.2012.02.003.

World Trade Organization (1994). Agreement on Implementation of Article VII of the General Agreement on Tariffs and Trade 1994 (Customs Valuation). In: Official Journal of the European Communities, L336, 23/12/1994, p. 119.

Yohai, V. J. (1987). High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, **15**, 642–656.

A. The mathematical ground of the "Fair Price" term

World Trade Organization (1994), the legal basis of international trade, states in its preamble and guiding principles the need to establish "a *fair*, uniform and neutral system for the customs valuation that precludes the use of arbitrary or fictitious customs values". The concept of "fairness" implies the respect of non-discrimination among the WTO partners, meaning that the valuation procedures enforced by a country must ensure to all WTO member nations equal trading conditions. For this reason, the agreement stipulates precise customs valuation rules aimed at avoiding discretion and enforcement of arbitrary criteria.

In this context, it becomes controversial to associate the term "Fair Price" to a statistical procedure that computes a baseline for customs valuation and post-clearance controls. Here we will not solve the controversy nor elaborate on this delicate issue. The aim of the section is to illustrate the mathematical motivations for the adoption of the "Fair Price" term, in this report, in Arsenis *et al.* (2015) and in THESEUS. To this end, we find appropriate to formulate the Fair Price in terms of *mathematical expectation*.

For the discussion, we use a subset of a COMEXT dataset concerning the import of a seafood from Canada into the EU (Riani *et al.*, 2008; Perrotta and Torti, 2010). These data contain some anomalous import flows, which appear under the fitted line in Figure 6: these flows are simply neglected in what follows (say that they are somehow detected and removed). We warn the reader that we will make some abuse of mathematical notation, to not make the reading to non-specialists cumbersome.

The linear regression model used to fit the good data is $y = \beta_0 + x\beta_1 + \epsilon$, where y is the n-vector of responses, x is a n-vector of known constants, and β_0 and β_1 are unknown parameters. Sometimes in COMEXT data we observe positive trade value flows associated to 0-quantities; this justifies to generalize in this section the discussion to a model with intercept. We want to link this statistical model with an equivalent probabilistic linear model, where for a fixed quantity traded x there is *uncertainty* in the traded value y. The uncertainty is determined by the *random variable* ϵ, which makes also y a random variable[11]. In other words, for a fixed quantity x, the observed values of y will differ from the expected value by a random amount ϵ.

Now, the trade flows $(x_1, y_1), \cdot, (x_n, y_n)$ found in COMEXT data for our seafood commodity are clearly scattered around the estimated regression line in Figure 6, but we do not know if the estimated line is representative of the *true* regression line. To analyze the problem, we can imagine to dispose of the universal

[11] This is the first major abuse of notation and terminology. A random variable is a function, defined in our case on the totality of possible trades (denoted as a set Ω), which associates each trade instance to a trade value expressed as real number (in the counterdomain $\Re$). The function should have been denoted as capital letter Y, to distinguish it from a specific value (realization) y.

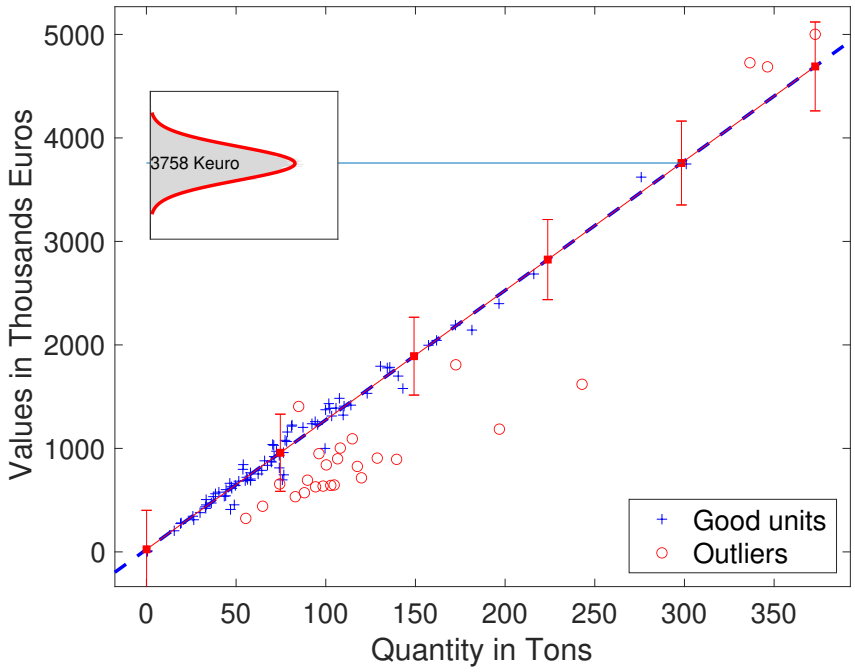

Figure 6: Fair Price illustration

trade of this seafood so that to compute $\mu_{y|x^*}$, the mean of all the possible values y that are traded at a certain fixed quantity $x = x^*$; similarly, we can indicate with $\sigma^2_{y|x^*}$ the variance of the same population, which measures how much the traded values y are spread around the mean value $\mu_{y|x^*}$. Figure 6 represents the spread as vertical red lines, which are noticeably identical: this is because we assume that $\sigma^2_{y|x^*}$ remains the same independently from x^* (*homoscedasticity* assumption).

If the random variable ϵ is assigned with a normal distribution with 0-mean, then the vertical lines are associated with a normal curve, each one with the same standard deviation (the square root of the variance), which determines the extent to which each curve spreads around the regression line. Real scenarios are maybe not so symmetric and smooth as the normal curve, but the point we want to make here is that *for each given quantity x, the probability to observe a trade value near the line is higher than to observe it far from it*. In this sense, the normal curve is telling a lot about how representative the linear fit actually is, because the closer the points are to the line and to the mode/mean/median of the normal distributions, the larger is their probability to occur. Figure 6 represents for illustration the normal spread around an *expected* value of 3758 K€for a trade of 298.3 tons of seafood.

Here is where the mathematical expectation and fair price come into play. The *expected trade value y* is an average over all the possible values y that one can observe, each value being weighted by the probability to observe it: it is a "fair value" for the uncertainty with which we are concerned when we have to take a decision about the regularity of a customs value. The same reasoning of course applies to the fair price, which is derived dividing the fair value by the trade quantity x.

From this perspective the WTO rules requesting that, "when there is more than one transaction value of same [or similar] goods, the lowest value will be used as the customs value", are implicitly adopting probabilities of Dirac-type: all values observed above the lowest value have probability 1, and those below have probability 0. The "fairness" concept here is applied having in mind the worst case rather than the average case scenario.

We conclude with few words on the way the linear model is computing our mathematical expectation. The intercept of the true regression line β_0 is the average value y when the quantity x is zero; the slope β_1 is the expected change in the traded values y associated to a unit increase of the traded quantity x of, say, one ton. To estimate conveniently these parameters from the data, we need another assumption. The trade flows observed in COMEXT are aggregates built from single customs declarations in a time period of a month. It is reasonable to assume that these trade flows are essentially independent one from the others. This assumption allows using the principle of least squares of Gauss to derive estimates for β_0 and β_1, which we indicate usually with $\hat{\beta}_0$ and $\hat{\beta}_1$: these estimates minimize the sum of the squared vertical deviations of the data from the line. Then these estimated parameters allow to *predict* the trade value for given quantity x^* as $\hat{y}^* = \hat{\beta}_0 + x^*\hat{\beta}_1$. The prediction is the trade value that we would *expect*, on the basis of the estimated regression, in association to the trade quantity $x = x^*$ or, equivalently, it is the estimated mean for this particular trade population and traded quantity x^*. The associated residual $y^* - \hat{y}^* = \epsilon^*$ is an estimate for the true error whose standard deviation is σ, which is a key statistic used for determining confidence intervals, hypothesis testing and other purposes which are beyond the scope of this section.

B. The adaptive fitting mechanism of the Forward Search

This section illustrates with a simple example the adaptive strategy used by the Forward Search (FS) for choosing the percentage of data units considered for fitting the model. We contrast the FS with the Least Trimmed Squares (LTS), which uses a percentage of data units h fixed in advance (actually the method offers a form of re-weighting that slightly increases the initial h, which is not considered here).

For the illustration, we use a COMEXT dataset already discussed in relation to the FS by Riani *et al.* (2008); Perrotta and Torti (2010), and more recently in the Bayesian context by Riani *et al.* (2018). The dataset concerns the import of a specific seafood from Canada into the EU. These data contain anomalous flows with smaller unit prices generated by fraudulent imports into one Member State (the outliers appear clearly under the fitted lines in both panels of Figure 7).

The left panel of the figure shows the estimated price obtained with the LTS. Here the data points used for the fit are a fixed 70% of the complete dataset, represented in black. The rest of the points (blue crosses) are not used for the fit, but of course not all of them are considered outliers. In other words, more data points could have been used for the price estimation. The right panel of the same figure shows the estimated price obtained with the FS. Here the data points used for the fit are found with an adaptive data-driven mechanism. The result is that now only the most deviating data points are left out from the fit (the black points are many more). In other words, with the FS almost all data points not used for the price estimation could be declared price outliers and investigated. Therefore, although the two fits do not differ much, 13.27 euro (LTS) vs 12.97 euro (FS), the FS estimate should be preferred.

Finally, Figure 8 shows how the prediction bands for the FS fit are used to identify the price outliers, which correspond precisely to the flows declared by the Member State where the undervaluation fraud occurred. In this case the bands obtained starting from the LTS fit produce comparable results.

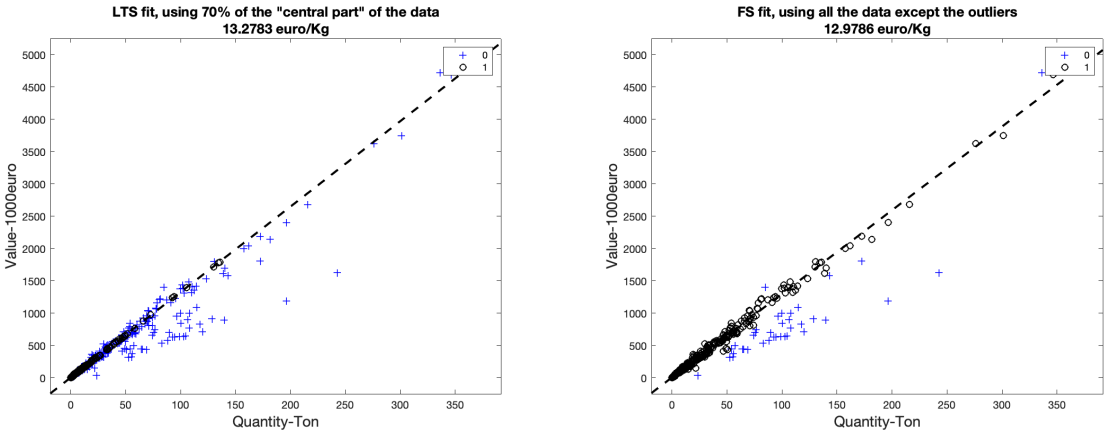

Figure 7: Right panel: the adaptive FS mechanism includes as much data as possible; almost all excluded points are clear price outliers; left panel: LTS fit based on a fixed (70%) percentage.

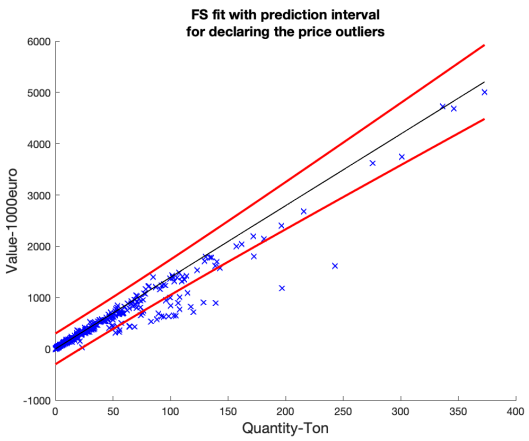

Figure 8: The prediction bands around the FS fit.

C. Bayesian approach in linear regression

In the linear regression model $y = X\beta + \epsilon$, y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row $x_i^\top$, and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. Bayesian approach in statistics allows to introduce in the estimate of β some prior knowledge. In a regression context, this generally reduces to assuming a prior distribution for the p–vector of coefficients β and a prior distribution for the regression standard error σ. The usual assumptions are:

$$\beta \sim N(\beta_0, s_0^2 R_0^{-1}) \text{ and } \sigma^2 \sim IG(a_0, b_0)$$

where IG stands for Inverse Gamma; $a_0 = \nu_0/2$; $b_0 = s_0^2\nu_0/2$ and ν_0 represents the degrees of freedom of the prior information. Actually, it is more common to assume that $\tau = 1/\sigma^2$ has a Gamma $G(a_0, b_0)$ prior distribution.

Using this prior information in the regression of a dependent variable y with respect to regressors X, the posterior distribution of β conditional on τ is $N\{\hat{\beta}_1, (1/\tau)(R_0 + X^\top X)^{-1}\}$ where:

$$\begin{aligned} \hat{\beta}_1 &= (R_0 + X^\top X)^{-1}(R_0\beta_0 + X^\top y) \\ &= (R_0 + X^\top X)^{-1}(R_0\beta_0 + X^\top X\hat{\beta}) \\ &= (I - A)\beta_0 + A\hat{\beta}, \end{aligned} \tag{C1}$$

and $A = (R_0 + X^\top X)^{-1}X^\top X$. The last expression shows that the posterior estimate $\hat{\beta}_1$ is a matrix weighted average of the prior mean β_0 and the classical OLS estimate $\hat{\beta}$, with weights $I - A$ and A. If prior information is strong, the elements of R_0 will be large, and A will be small, so that the posterior mean gives most weight to the prior mean[12].

The posterior distribution of τ is $G(a_1, b_1)$ where

$$a_1 = a + n/2 = (\nu_0 + n)/2 \quad \text{and} \tag{C2}$$

$$b_1 = \left\{\nu_0/\tau_0 + (y - X\beta_1)^\top y + (\beta_0 - \beta_1)^\top R_0\beta_0\right\}/2. \tag{C3}$$

The posterior distribution of σ^2 is $IG(a_1, b_1)$. The posterior mean estimates of τ and σ^2 are respectively

$$\tau_1 = a_1/b_1, \quad \text{and} \quad \tilde{\sigma}_1^2 = b_1/(a_1 - 1). \tag{C4}$$

Unless a_1 is very small, the difference between $\hat{\sigma}_1^2$ and $\tilde{\sigma}_1^2$ is negligible.

[12] In the classical Bayesian approach these weights are fixed, while with the Bayesian Forward Search, as the subset of current observations considered for the fit grows, the weight assigned to A increases. So we can dynamically see how the estimate changes as the effect of the prior decreases.

D. Prior information and previous observations

In fraud detection (see for example Perrotta and Torti (2010)), we have appreciable prior information about the values of the parameters. This can conveniently be thought of as coming from n_0 fictitious observations y_0 with matrix of explanatory variables X_0. Then the data consist of the n_0 fictitious observations plus n actual observations. Where available, the device of prior observations provides a convenient representation of prior information. In distributional terms it provides independent conjugate prior distributions for the parameters; normal for β and inverse gamma for σ^2. We follow, for example, Chaloner and Brant (1988) who are interested in outlier detection, and describe the parameter values of these prior distributions in terms of the fictitious observations of the previous section.

We start with σ^2. Let $\tau = 1/\sigma^2$. In the notation of C, the prior distribution of τ is gamma $G(a_0, b_0)$ and that of σ^2 is inverse gamma $IG(a_0, b_0)$. The estimate of σ^2 from the n_0 fictitious observations is s_0^2 with $\nu_0 = n_0 - p$ degrees of freedom. Then

$$a_0 = \nu_0/2 = (n_0 - p)/2 \quad \text{and} \quad b_0 = \nu_0 s_0^2/2 = s_0^2(n_0 - p)/2.$$

Prior information for the linear model is given as the scaled information matrix $R_0 = X_0^{\mathsf{T}} X_0$ and the prior mean $\hat{\beta}_0 = R_0^{-1} X_0^{\mathsf{T}} y_0$. Then $S_0 = y_0^{\mathsf{T}} y_0 - \hat{\beta}_0^{\mathsf{T}} R_0 \hat{\beta}_0$. Thus, given n_0 prior observations the parameters for the normal inverse-gamma prior may readily be calculated. This information is then combined with that provided by the n current observation for deriving the posterior distribution of β and σ, as described in C.

E. Bayesian search

In absence of prior information, the FS uses least squares to fit the model to subsets of m observations, starting from an initial subset of m_0 observations. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest squared residuals. For each m ($m_0 \leq m \leq n-1$), we test for the presence of outliers, using the observation outside the subset with the smallest absolute deletion residual. Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors and the vector of dependent variable are respectively $X(m)$ and $y(m)$. Ordinary least squares on this subset of observations yields parameter estimates $\hat{\beta}(m)$ and $\hat{\sigma}^2(m)$, an estimate of σ^2 on $m-p$ degrees of freedom. The residuals for all n observations, including those not in $S^*(m)$, are

$$e_i(m) = y_i - x_i^{\mathsf{T}}\hat{\beta}(m) \qquad (i = 1, \ldots, n).$$

To test for outliers, the deletion residuals are calculated for the $n-m$ observations not in $S^*(m)$. These residuals are:

$$r_i(m) = \frac{y_i - x_i^{\mathsf{T}}\hat{\beta}(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_i(m)\}}} = \frac{e_i(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_i(m)\}}},$$

where $h_i(m)$ represents the leverage of x_i, that is: $h_i(m) = x_i^{\mathsf{T}}\{X(m)^{\mathsf{T}}X(m)\}^{-1}x_i$. Let the observation nearest to those forming $S^*(m)$ be $i_{\min}$ where

$$i_{\min} = \arg\min_{i \notin S^*(m)} |r_i(m)|.$$

To test whether observation i_{min} is an outlier we use the absolute value of the minimum deletion residual

$$r_{i_{min}}(m) = \frac{e_{i_{min}}(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_{i_{min}}(m)\}}},$$

as a test statistic. If the absolute value of $r_{i_{min}}(m)$ is too large, the observation i_{min} is considered to be an outlier, as well as all other observations not in $S^*(m)$.

In order to test for outliers we need a reference distribution for the residuals $r_i(m)$. If we estimated σ^2 from all n observations, the statistics would have a t distribution on $n-p$ degrees of freedom. However, in the search we select the central m out of n observations to provide the estimate $s^2(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, let the variance of the symmetrically truncated standard normal distribution containing the central m/n portion of the full distribution be

$$c(m,n) = 1 - \frac{2n}{m}\Phi^{-1}\left(\frac{n+m}{2n}\right)\phi\left\{\Phi^{-1}\left(\frac{n+m}{2n}\right)\right\},$$

where $\phi(.)$ and $\Phi(.)$ are respectively the standard normal density and c.d.f. See Riani *et al.* (2009) for a derivation from the general method of Tallis (1963). We take $s^2(m)/c(m,n)$ as our approximately unbiased estimate of variance. In the robustness literature, the important quantity $c(m,n)$ is called a consistency factor (Riani *et al.*, 2014a; Johansen and Nielsen, 2015).

The Bayesian approach with prior information based on n_0 ficticious (previous) observations gives us a robust starting point for the search. So, in this case, the search can proceed from $m=0$ (i.e. the fictitious observations provide the parameter values β_0, a_0, b_0, ν_0 and R_0 for all n residuals from the data). The procedure then continues as outlined above except that the n_0 prior observations are always included in the search; their residuals are ignored in the selection of successive subsets. The algebra for the FS with prior information is similar to that of the frequentist search, but there is one complication. The n_0 fictitious observations are treated as a sample with population variance σ^2. The m observations from the actual data are from a truncated distribution of m out of n observations and so asymptotically have a variance $c(m,n)\sigma^2$. An adjustment must be made before the two samples are combined. This becomes a standard problem in weighted least squares (for example, Rao 1973, p. 230). Let y^+ be the $(n_0+m)\times 1$ vector of responses from the fictitious observations and the subset, with X^+ the corresponding matrix of explanatory variables. The covariance matrix of the independent observations is $\sigma^2 G$, with G a diagonal matrix; the first n_0 elements of the diagonal of G equal one and the last m elements have the value $c(m,n)$. The information matrix for the n_0+m observations is:

$$(X^{+\mathsf{T}}WX^+)/\sigma^2 = \{X_0^{\mathsf{T}}X_0 + X(m)^{\mathsf{T}}X(m)/c(m,n)\}/\sigma^2, \tag{E5}$$

where $W = G^{-1}$. In the least squares calculations we need only to multiply the elements of the sample values of y and X by $c(m,n)^{-1/2}$. This implies that, at step m of the search, the posterior estimate of β is

given by:

$$\begin{aligned}\hat{\beta}_1(m) &= (X^{+\mathsf{T}}WX^{+})^{-1}X^{+\mathsf{T}}Wy^{+} \\ &= \{X_0^{\mathsf{T}}X_0 + X(m)^{\mathsf{T}}X(m)/c(m,n)\}^{-1}\{X_0^{\mathsf{T}}y_0 + X(m)^{\mathsf{T}}y(m)/c(m,n)\} \\ &= \{X_0^{\mathsf{T}}X_0 + X(m)^{\mathsf{T}}X(m)/c(m,n)\}^{-1}\{X_0^{\mathsf{T}}X_0\hat{\beta}_0 + X(m)^{\mathsf{T}}y(m)/c(m,n)\}\end{aligned}$$

whereas the values of the leverages are given by

$$h_i(m) = x_i^{\mathsf{T}}\{X_0^{\mathsf{T}}X_0 + X(m)^{\mathsf{T}}X(m)/c(m,n)\}^{-1}x_i.$$

More details on this issue can be found in Atkinson *et al.* (2016a) and Atkinson *et al.* (2017).

F. The effect of prior information on envelopes and model parameters

A Bayesian FS through the data provides a set of n absolute minimum deletion residuals. We require the null pointwise distribution of this set of values and find, for each value of m, a numerical estimate of, for example, the 99% quantile of the distribution of $|r_{i_{min}}(m)|$. When used as the boundary of critical regions for outlier testing, these envelopes have a pointwise size of 1%.

We now illustrate the effect of prior information on the envelopes. Figure 9 shows the results of 10,000 simulations of normally distributed observations from a regression model with four variables and a constant ($p = 5$), the values of the explanatory variables having independent standard normal distributions. These envelopes are invariant to the numerical values of β and σ^2. The left-hand panel shows 1, 50 and 99% simulation envelopes for weak prior information when $n_0 = 30$ (and $n = 500$), along with the envelopes in the absence of any prior information. As m increases the two sets of envelopes become virtually indistinguishable, illustrating the irrelevance of this amount of prior information for such large samples. On the other hand, the right-hand panel keeps $n = 500$, but now n_0 has the same value. There is again good agreement between the two sets of envelopes towards the end of the search, especially for the upper envelope.

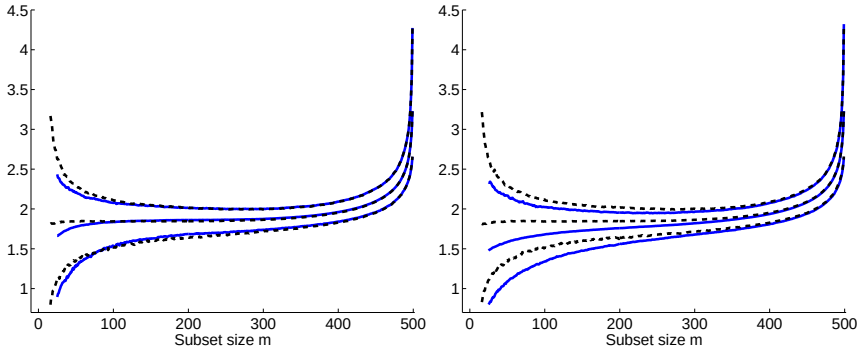

Figure 9: The effect of correct prior information on forward plots of envelopes of absolute Bayesian minimum deletion residuals. Left-hand panel, weak prior information ($n_0 = 30; n = 500$). Right-hand panel, strong prior information ($n_0 = 500; n = 500$), 10,000 simulations; 1, 50 and 99% empirical quantiles. Dashed lines, without prior information; heavy lines, with prior information

The effect of prior information on the parameter estimation is instead represented in Figure 10. The left-hand panel of the Figure shows empirical quantiles for the distribution of $\hat{\beta}_3(m)$ from the 10,000 simulations when $\beta_3 = 0$. Because of the symmetry of our simulations, this is indistinguishable from the plots for the other parameters of the linear model. The right-hand panel shows the forward plot of $\hat{\sigma}^2(m)$, simulated with $\sigma^2 = 1$. In this simulation the prior information, with $n_0 = 30$, is small compared with the sample information. In the forward plot for $\hat{\beta}_3$ the bands are initially wide, but rapidly narrow, being symmetrical about the simulation value of zero. There are two effects causing the initial rapid decrease in the width of the interval during the FS. The first is under-estimation of σ^2 which, as the right-hand panel shows, has a minimum value around 0.73. This under-estimation occurs because $c(m, n)$ is an asymptotic correction factor. Further correction is needed in finite samples. The second effect is again connected with the value of $c(m, n)$, which is small for small m/n (for example 0.00525 for 10%). Then, from (E5), the earliest observations to enter the search will have a strong effect on reducing var $\hat{\beta}(m)$.

The panels of Figure 11 are for similar simulations, but now with n_0 and n both 500. The main differences from Figure 10 are that the widths of the bands now decrease only slightly with m and that the estimate of σ^2 is relatively close to one throughout the search.

The incorporation of correct prior information into the analysis of data leads to parameter estimates with higher precision than those based just on the sample. There is a consequential increase in the power of tests about the values of the parameters and in the detection of outliers. Figure 12 shows average power curves[13] for Bayesian and frequentist procedures and also for Bayesian procedures with incorrectly specified priors when the contamination rate is 5% and $n_0 = 500$. The curves do not cross for powers a little less than 0.2 and above. The procedure with highest power is the curve that is furthest to the left which, in the figure, is the correctly specified Bayesian procedure. The next best is the frequentist one, ignoring prior information. The central power curve is that in which the mean of β_0 is wrongly specified as -1.5. This is the most powerful procedure for small shifts, as the incorrect prior is in the opposite direction to positive

[13] Average power is defined as the average proportion of contaminated observations correctly identified.

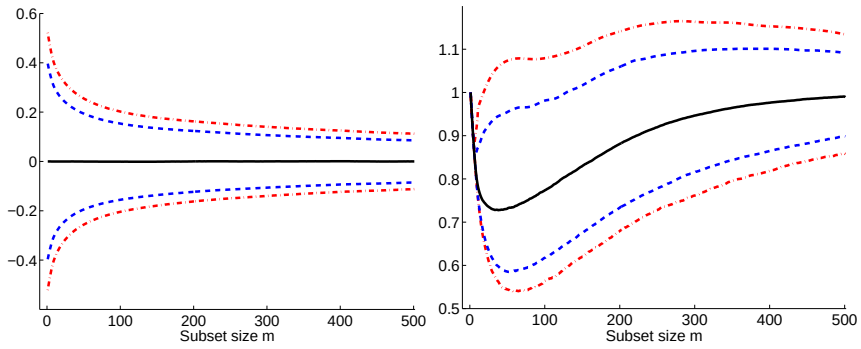

Figure 10: Distribution of parameter estimates when $\beta_3 = 0$ and $\sigma^2 = 1$. Left-hand panel $\hat{\beta}_3(m)$, right-hand panel $\hat{\sigma}^2(m)$; weak prior information ($n_0 = 30$; $n = 500$). 1, 5, 50, 95 and 99% empirical quantiles

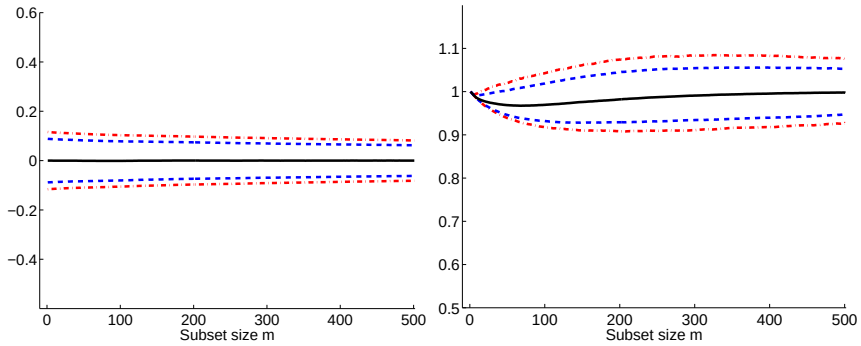

Figure 11: Distribution of parameter estimates when $\beta_3 = 0$ and $\sigma^2 = 1$. Left-hand panel $\hat{\beta}_3(m)$, right-hand panel $\hat{\sigma}^2(m)$; weak prior information ($n_0 = 500$; $n = 500$). 1, 5, 50, 95 and 99% empirical quantiles

quantity used to generate outliers. With large shifts, this effect becomes less important. For most values of average power, the curve for mis-specified σ^2 comes next, with positive mis-specification of β worst. Over these values, three of the four best procedures have power curves which are virtually translated horizontally. However, the curve for mis-specified β has a rather different shape at the lower end. With β mis-specified, the envelopes for large m sometimes lie slightly above the frequentist envelopes. The effect is to give occasional indication of outliers for relatively small values of the shift generating the outliers.

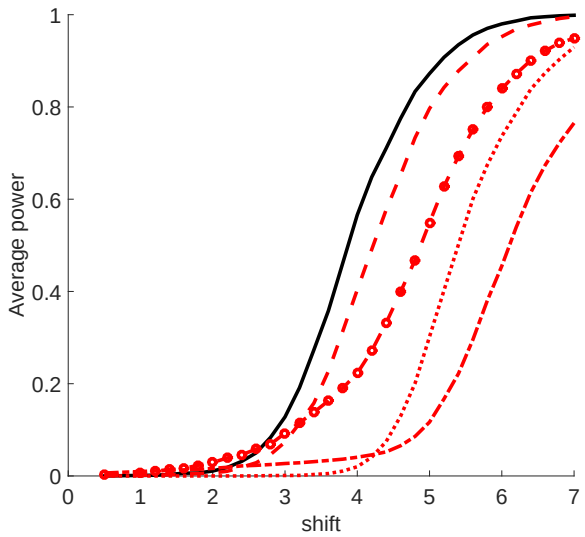

Figure 12: Average power in the presence and absence of prior information: $\sigma^2 = 1$. Reading across at a power of 0.6: Bayesian, solid line; frequentist, dashed line; wrong $\beta_0 = -1.5$, dashed line with circles; wrong $\sigma_0^2 = 3$, dotted line; wrong $\beta_0 = 1.5$, dotted and dashed line. Contamination 5%, 2,000 simulations, strong prior information; $n_0 = 500$.

In Figure 13, for 30% contamination, the Bayesian procedure is appreciably more powerful than the frequentist one, which is slightly less powerful than that with mis-specified σ_0^2. The rule for mis-specified $\beta_0 = 1.5$, has the lowest power, appreciably less than that in which $\beta_0 = -1.5$. Although the curves cross over for shifts around 3.5, the Bayesian procedure with correctly specified prior has the best performance until the shift is sufficiently small that the power is negligible.

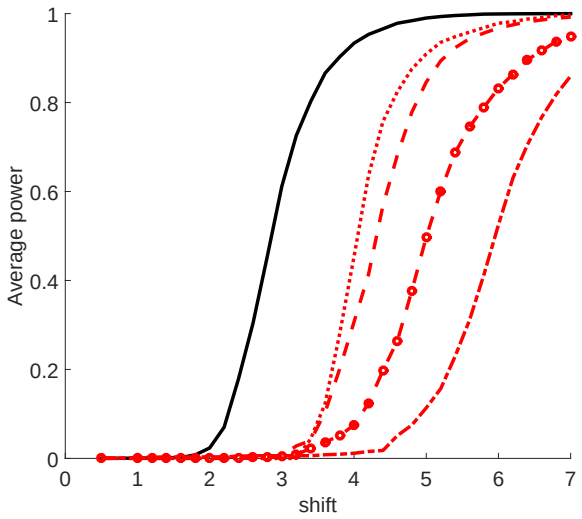

Figure 13: Average power in the presence and absence of prior information: $\sigma^2 = 1$. Reading across at a power of 0.6: Bayesian, solid line; wrong $\sigma_0^2 = 3$, dotted line; frequentist, dashed line; wrong $\beta_0 = -1.5$, dashed line with circles; wrong $\beta_0 = 1.5$, dotted and dashed line. Contamination 30%, 2,000 simulations, strong prior information; $n_0 = 500$.

G. R^2 correction

The value of the R^2 provides information about the goodness of fit of the regression. A value too low for the R^2 is an indication about lack of fit of the model (due for example to the presence of multiple populations or to heteroskedasticity). Similarly, a too high value of R^2 indicates the presence of a (almost) perfect fit and may lead to declare as outliers units which differ only slightly from the regression line constructed using the bulk of the data. The problem of perfect fit has been discussed also in Maronna *et al.* (2006). Figure 14 shows a typical perfect fit situation that often characterizes international trade data. Here the 21 observations lay almost on a straight line and the FS identifies 6 outliers. Most of them are however very close to the regression line. The very large value of the R^2 suggests that the number of signals may depend on a perfect fit problem.

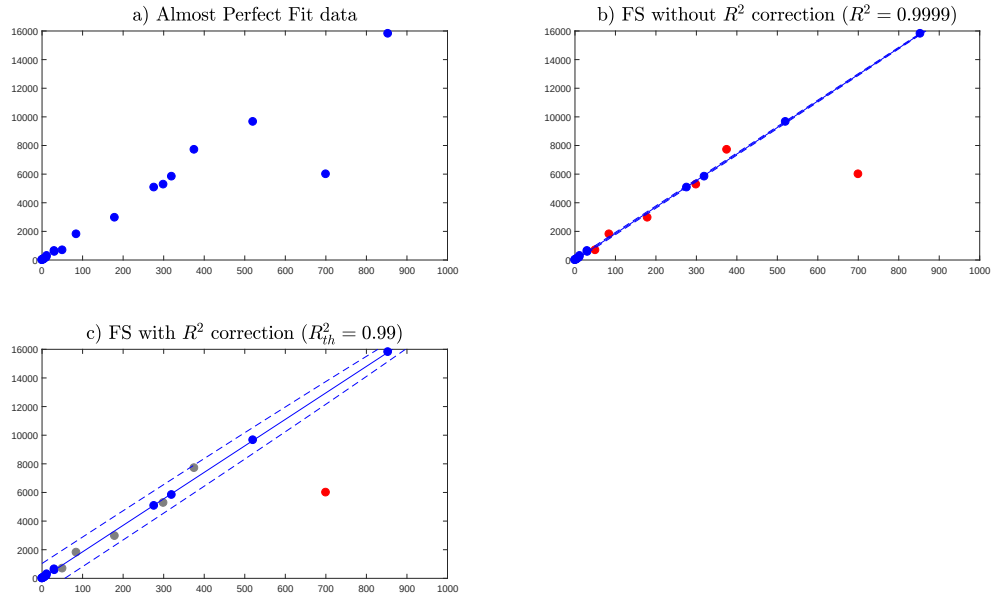

Figure 14: The application of R^2 correction on data with (almost) perfect fit

The idea for avoiding the spurious over-decaration of signals is to correct the mean square error according to a pre-determined maximum threshold R^2_{th}. So, when the R^2 of the regression is larger than the threshold, the correction a new mean square error is calculated as:

$$MSE_{th} = \frac{1 - R^2_{th}}{1 - R^2} MSE \tag{G6}$$

Since $R^2 > R^2_{th}$, then $\frac{1-R^2_{th}}{1-R^2} > 1$ which, in turn, implies that $MSE_{th} > MSE$. Therefore the corrected bands for the studentised residuals are larger than the original, and the number of observations outside the bands is expected to decrease. This effect is clearly highlighted in panel c of Figure 14: by imposing a threshold of 0.99 we can enlarge the studentised residuals confidence bands and sensibly reduce the number of identified outliers. In particular, after the correction, only 1 signal remains valid, whereas the others, being closer to the regression line, are not anymore considered anomalous.

For our estimates in Theseus we set $R^2_{th} = 0.95$.

H. The choice of n_0

The robust prices published in the "Monthly Fair Price" section of Theseus are calculated using the FS based on prior information described in the previous section. In particular, past monthly transactions on the same product P and origin O are used as a prior for estimating the price of next month. The main idea is that what happened on the market during the past is a good basis for analyzing the current situation. The choice of n_0, that is the number of past observations to consider for estimating the current price, must then guarantee a good balance between providing a stable representation of the market and allowing, at the same time, the natural evolution of price dynamics. In fact, being the posterior price a weighted average between past and current observations (see equation C1), using a value for $n_0 >> n$ would likely mask typical price movements, such as, for example, level shifts or seasonal cycles. On the contrary, choosing a too small value for n_0 may cause spurious volatility. Therefore, the rule for choosing n_0 must take into account the number of current observations available each month.

Once fixed the product P, the origin O and the month T, the maximum number of current observations is 28 (one for each MS). Several empirical applications proved that a good trade–off between stability of the estimates and dynamic representation was offered by the following rule:

$$n_0(T) = \begin{cases} 10, & \text{if } n(T) < 5 \\ 2n(T), & \text{if } 5 \leq n(T) < 20 \\ 40, & \text{if } n(T) \geq 20. \end{cases} \tag{H7}$$

Obviously, in the construction of the prior sample made of n_0 past transactions, only observations that contributed to the estimations of past prices are considered. Moreover, precedence is given to observations of month $(T-1)$, then the ones of month $(T-2)$ and so on, until the desired sample size is reached.

I. A practical example

Figure 15 shows how the issues illustrated in Sections F, G and H are reflected in the data and the published results. For this particular PO*T combination, we have only 6 observations, one of which is not used for the estimation because its quantity is 0 (zero x-values do not change the estimated slope of a regression without intercept). The prior (past) observations are then crucial to obtain a reliable estimate of the price. According to expression (H7), we select 10 past observations and use them to robustly estimate the regression line. A signal is detected (represented with a red cross), whereas the observation represented by a gray diamond (i) is not identified as outlier because of the R^2 correction but (ii) does not contribute to the robust fit as explained in Section G.

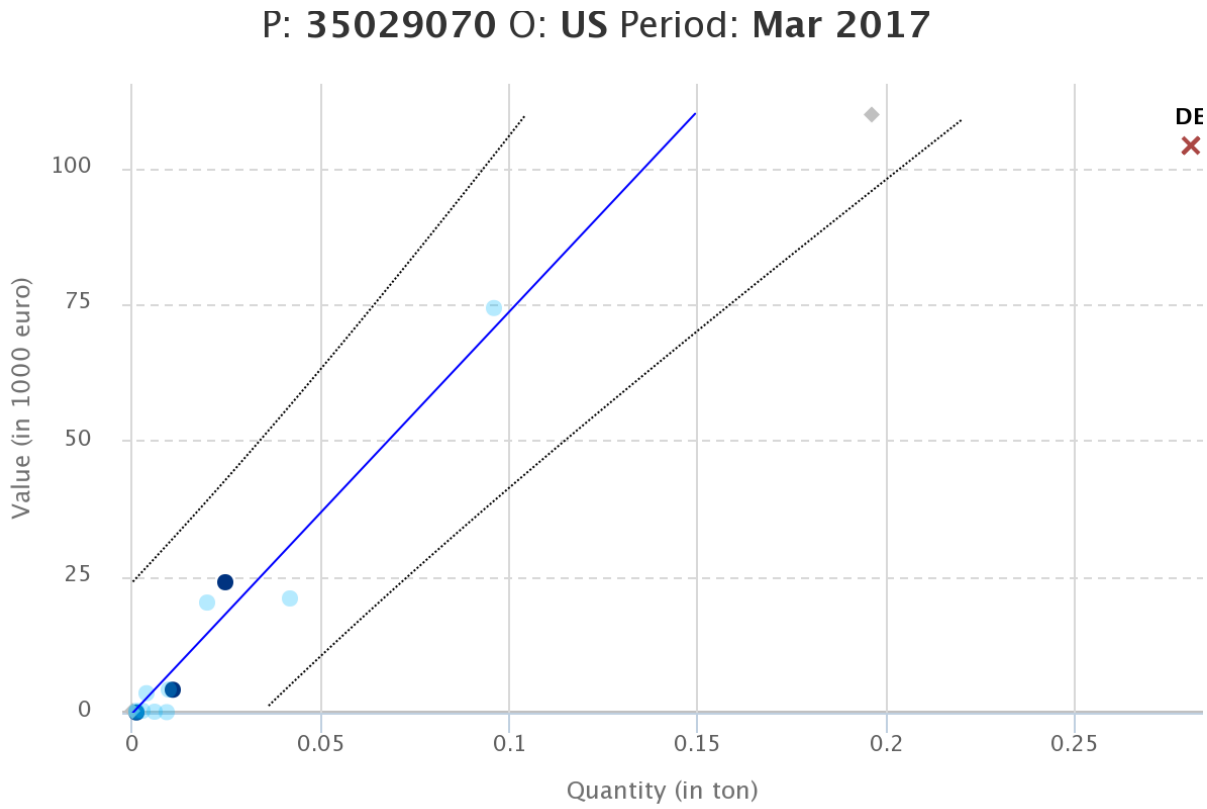

Figure 15: Forward search, prior observations and R^2 correction in a practical example

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-30188-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:10.2760/635844

ISBN 978-92-76-18351-8