Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

27-Oct-2011

English - Or. English

STATISTICS DIRECTORATE

STD/TBS/WPTGS(2011)7
Unclassified

Working Party on International Trade in Goods and Trade in Services Statistics

USING CLUSTER ANALYSIS FOR IDENTIFYING OUTLIERS AND POSSIBILITIES OFFERED WHEN CALCULATING UNIT VALUE INDICES

7-9 November 2011, OECD Headquarters, Paris

This paper is for discussion and was prepared by Evangelos Pongas for item 5.5 of the agenda

Contact person: Evangelos PONGAS, E-mail: Evangelos.Pongas@ec.europa.eu

JT03310111

English - Or. English

USING CLUSTER ANALYSIS FOR IDENTIFYING OUTLIERS AND POSSIBILITIES OFFERED WHEN CALCULATING UNIT VALUE INDICES

(CLUSTERING METHODS, HIDIROGLOU AND BERTHELOT, MAD ON MAIN PARAMETERS OF THE DISTRIBUTION OF DETAILED DATA)

EUROSTAT – G5
Evangelos Pongas, Aura Leulescu

Introduction

1. Trade Statistics are supplied monthly to Eurostat. Often they contain not only the last period but also revisions for previous periods. Data blocks (imports, exports by trade type: intra, extra) contain important errors mainly due to modifications of systems at Member State level. Such errors are not detected at national level because national publication and dissemination are based on outputs done at an earlier stage (than sending to Eurostat) of the production process. G5 is currently developing a validation package to control the data just before sending. The package validations are defined in a validation framework that has been agreed with Member States.

2. Error detection at Eurostat is done at two levels:

Primary validation:

- Before loading to Comext, data are controlled and transformed with XT-NET Edit software. The process is streamlined and automated to permit the fast management and loading of many huge datasets. This production process is more focused on data transformation and assessment of consistency with main classification (aggregations etc.) rather than on validation (only basic variations are done). The system is rather rigid and evolution of validations is a heavy process.

Post validation:

- - Once per month, G5 checks time series of total trade by individual partner. Since control is based on totals, detailed data problems are not detected. Detailed data outlier detection is also possible, but correction process is too lengthy and limited to major errors.
- - G4 runs systematically mirror outlier detection but the error correction, as in the previous case is mostly focused on major problems and has problems and delays.
- - Users often detect more detailed problems and communicate them to G5.

3. The aim of this paper is to give an overview of the current development work that unit G5 is undergoing for the validation of the external trade data. These methods are experimental, but they might be used already from 2012 in the regular production process.

METHOD 1: DETECTION OF IMPORTANT INFLUENTIAL DATA ERRORS IN INTERNATIONAL TRADE OF GOODS STATISTICS

Objective

4. To define a statistical approach that will permit a macro "important error" detection satisfying the following criteria.

- - Fast program execution (all countries, many periods in less than ½ hour)
- - Easy and clear decision making support (keep or erase wrong data in Comext)
- - Orientations for fast and easy problem localisation.

5. The current proposal intends to fill the existing gaps in the current validation system and to give some insurance that globally the published statistics are of acceptable quality.

6. Problems detected by the method refer to important influential (having impact on published aggregates) errors and not individual errors that have no impact on aggregated figures.

Method

7. The proposed test is based on the assumption that detailed data trade distributions are stable in short time terms (2-3 years). Thus big changes of the main distribution of monthly aggregates (MA) (sum, average, standard deviation, count, maximum) are probably due to errors. The list of these five aggregates could be extended to include other moments like skewness and kurtosis. Since standard SQL does support such functions, program development effort and execution time would increase considerably.

8. The test is based on two main operations:

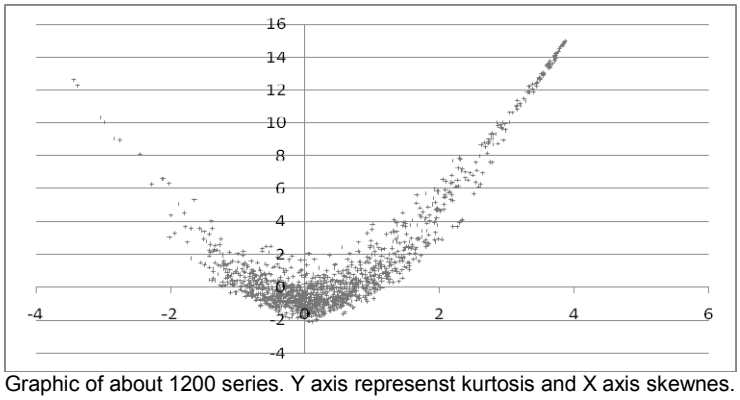
- Calculate the five previous main aggregates for each month of a period of not less than 13 and not more than 36 periods. The calculation can be done with a single SQL (see example at the end of the document) and it takes around ten minutes per year for all EU countries together. Join the results of the annual queries in order to compose five time series (variables) per county/flow/trade type. In total, 540 (27*4*5) times series are constructed. Each time series contains 13 – 36 figures.

Normalise all time series (Ym) and construct new variables (Zm) with the following function:

Zm= (Ym – average(Y)) / standard deviation (Y) where m=1...number of periods. The new variables Z are (approximately) normally distributed with mean zero and standard deviation one: N(0,1). Zm are equivalent to z-scores = how many times they are far in terms of standard deviation from the centre of the distribution (mean). For the N(0,1) distribution, 99.7 of Zm are less than 3 (or more than -3), 95% than 2 and 68% than 1. It is known as the rule 68, 95, 99.7

Therefore if one or more Zm are >3 or <-3, the probability to have severe errors is extremely high and investigations must be done immediately. If all five Zm for a given flow/trade type are >3 or <-3, it is preferable to revert immediately in Comext the data month to the previous state and then investigate. Some examples at the end of the paper show in real terms the error detection and possible explanations. Often trade distributions are right skewed. In this case different left and right thresholds can be applied (example -2.5 and 3).

Kurtosis and skewness of Y series (Z gives same results) can be used for a global evaluation of the health of ITG statistics. Kurtosis and skewness have a quadratic relationship among them for the case of ITGS. Each series is represented by a single point. Problematic series are at the extremes of the curve. Left extremes indicate a probable under-reporting while right extremes indicate a probable over-reporting.



Graphic of about 1200 series. Y axis represenst kurtosis and X axis skewnes.

9. The above approach is based in traditional statistics and measures. If many periods are erroneous, the traditional statistical results are contaminated by the errors. It is therefore proposed to use in parallel robust methods (based on median, robust variance, robust z-score). The combination of the two methods results will surely permit a better error detection and decision making.

10. The following examples show some test results of the method. Problems kept use the rule *z_score >3 or <-3*

DECLARANT	**CZ**				
TRADE_TYPE	**I**				
FLOW	**1**				
	period				
variable	**201011**	**201012**	**201101**	**201102**	**201103**
Avg_Q		3.54			
Avg_SQ		3.45			
Avg_Val		3.53			
Count_Q		-3.48			
Count_SQ		-3.48			
Count_Val		-3.48			
Max_Q		3.15			
Max_SQ				3.15	
Max_Val		3.6			
SUM_Q					
SUM_SQ					
SUM_Val					
StdDev_Q		3.56			
StdDev_SQ		3.14			
StdDev_Val		3.61			

11. It is obvious that Dec CZ imports contain major errors. Possible errors in supplementary quantities in Feb 2011 (error removed by CZ from latest data version).

DECLARANT	**GR**				
TRADE_TYPE	**E**				
FLOW	**2**				
	period				
variable	**201010**	**201011**	**201012**	**201101**	**201102**
Avg_Q					
Avg_SQ		3.41			
Avg_Val					
Count_Q					
Count_SQ					
Count_Val					
Max_Q					
Max_SQ		3.57			
Max_Val					
SUM_Q					
SUM_SQ		3.37			
SUM_Val					
StdDev_Q					
StdDev_SQ		3.56			
StdDev_Val					

12. Probable error in supplementary quantities in Nov 2011. The verification shows at least a problem in bold below:

DECLARANT	GR
FLOW	2
PRODUCT_TARIC	90283011
PARTNER	288
VALUE201008	27.925
QUANTITY201008	2472
SUP_QUANTITY201008	1400
VALUE201011	53.835
QUANTITY201011	4613
SUP_QUANTITY201011	**300015096**
VALUE201012	35.89
QUANTITY201012	3169
SUP_QUANTITY201012	2000

METHOD 2: DETECTION OF OUTLIERS WITH THE USE OF BERTHELOT AND HIDIROGLOU METHOD

Objective

13. The main objective of the method is to detect important (weighted) outliers in the primary data. This method is similar to the scored system applied by several countries

Method (Hidiroglou and Berthelot – ratio method (Garcia et al, 2006))

14. The HB edit uses price ratio, $p = V/Q$, and p2 , the median of unit prices. The unit price ratios are then transformed in two steps:

a) centring transformations on ratios

$$S_i = \begin{cases} p_i / p_2 - 1 \; if \;\; p_i \geq p_2 \\ 1 - p_2 / p_i \; if \;\; 0 < p_i \leq p_2 \end{cases},$$

b) magnitude transformation that accounts for the relative importance of large cases: $E_i = S_i * \{\max(V_i, p_2 * Q_i)\}^u$, where $0 < u < 1$.

c) A further transformation is needed to account for data that are highly clustered around the median:

$$d_{q_1} = \max(q_2 - q_1, abs(a * q_2))$$
$$d_{q_3} = \max(q_3 - q_2, abs(a * q_2))$$

15. .We assign to every observation a score that is a ratio with a factor measuring displacement of unit prices from the median, weighted by the appropriate distance from the median.

$$Ratio_i = \begin{cases} (q_2 - p_i) / d_{q_1} \; if \;\; p_i < q_2 \\ (p_i - q_2) / d_{q_3} \; if \;\; p_i > q_2 \end{cases}$$

16. This method solves several problems: it reduces skewness, works well on both sides of the tail, and accounts for the 'magnitude' of potential outliers. This allows us to account in the selection for the importance of deviations in large units in terms of value or quantity.

We use an impact function for selection that account for the importance of the error and the potential impact

We include another impact function based on the measure developed by Jäder and Norberg (2005) for Swedish trade data:

$$Diff_i = abs(V^r_{i,cm} - p_2 * Q^r_{i,cm}) / Total(V_{cm}) .$$

where P2*Qm are estimates of the expected value of shipments for unit *i* respectively. The **estimated total for each commodity** is calculated using final data for records accepted or automatically imputed by the automated system

- **Problem: due to the inherent nature of some products that have high variation we have a large number of type I errors-" false outliers". These products, such as wines or electronics are often stratified in low-end and high-end markets. For these cases we decided first to cluster and then to select outliers that fall out of these intervals.**
- **This univariate method is generally preferred to multivariate mainly due to its simplicity to use. However it cannot detect records that violate the correlation structure of the data.**
- **In order to treat these issues, we have also tried a clustering approach, the k-means clustering method. From the results of the tests we conclude the efficiency of the suggested methods compared to the application of simple univariate methods**.

METHOD 3: DETECTION OF OUTLIERS WITH THE USE K- MEANS CLUSTERING METHOD

Objective

17. The initial objective of the method was to complete the previous methods of outlier detection. However we consider that the clustering methods can also be used to:

- To detect inliers in the primary data or, eventually, misclassification of products.
- To detect automatically sub-products
- To define single or multiple validation intervals
- To clean the data.

Method

18. The first step is to construct blocks of data to submit for clustering. This can be done by grouping the data with the use of a sub key of the records composed by only some fields. In our experiment, we did the grouping by CN8 product, month, reporting country, trade flow and trade type (extrastat, intrastate). Mode of transport could be included since it might have an important impact on the value of imported goods. Since the data received by Eurostat are already aggregated, some outliers are masked. It is expected that application at national level with micro data will be more efficient.

19. The second step is the standardization the data. The data are normally in two or three vectors (value, net mass, supplementary quantity if available). The data might also be transformed to ratios before standardization (value/net-mass, value/supplementary quantity, net-mass/supplementary quantity)

20. The standardised vectors are clustered with the k-means method. Hierarchical clustering was also tested but abandoned due to high process time. For the same reason, even in the k-means method, if the length of arrays is big, then distances are calculated by sampling (not for each point to all others but for each point to a sample of others). In k-means method the number of clusters must be predefined. After some preliminary results we limited the number of clusters from two to five. However in many cases, the 3 clusters division gave better outlier results and the two clusters division gave better sub-products.

21. The approach allows to identify stratified products and to establish homogenous groups inside the CN8 level data. Thus we identify the outliers based on two criteria:

- The modification of R-square: it must be significant and the distance of the point from the centroid must be high.
- The **number of observations** in the new cluster: it must be very low so we can consider them outliers, rather than a cluster of differentiated products.

Total sum of squares TTS=Inter-class sum of squares + Intra-classes sum of squares

R-square=inter-classes sum of squares/TSS is the ratio of the variance explained through clustering and the total variance.

22. For the selection of the most important outliers we add a measure of displacement in terms of Euclidian distances.

$$fact\_dist = \frac{(dist - HL)}{av\_dist}$$, HL=av_dist+3,5*sdev_dist

In case of inliers, the displacement can not be applied and other criteria have to be found.

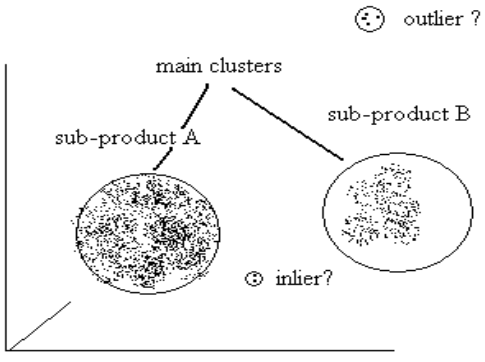
Clustering enables us to reduce the probability of type of errors observed with the Berthelot and Hidiroglou method

23. After simulations we decided to chose the number of clusters when the R-square is more than 50% . We skip to a higher number of clusters when the improvement of R-square is more than 10%. If R-square is less than 50% we consider than the clusters method is not efficient and we apply traditional statistical methods based on average and standard deviation. The same approach is kept if the number of observations in the block is small (<15)

24. In the future we will also test the Mahalanobis distance instead of Euclidian.

25. Having applied the clusters method and having eliminated or corrected the outliers, further process can be introduced:

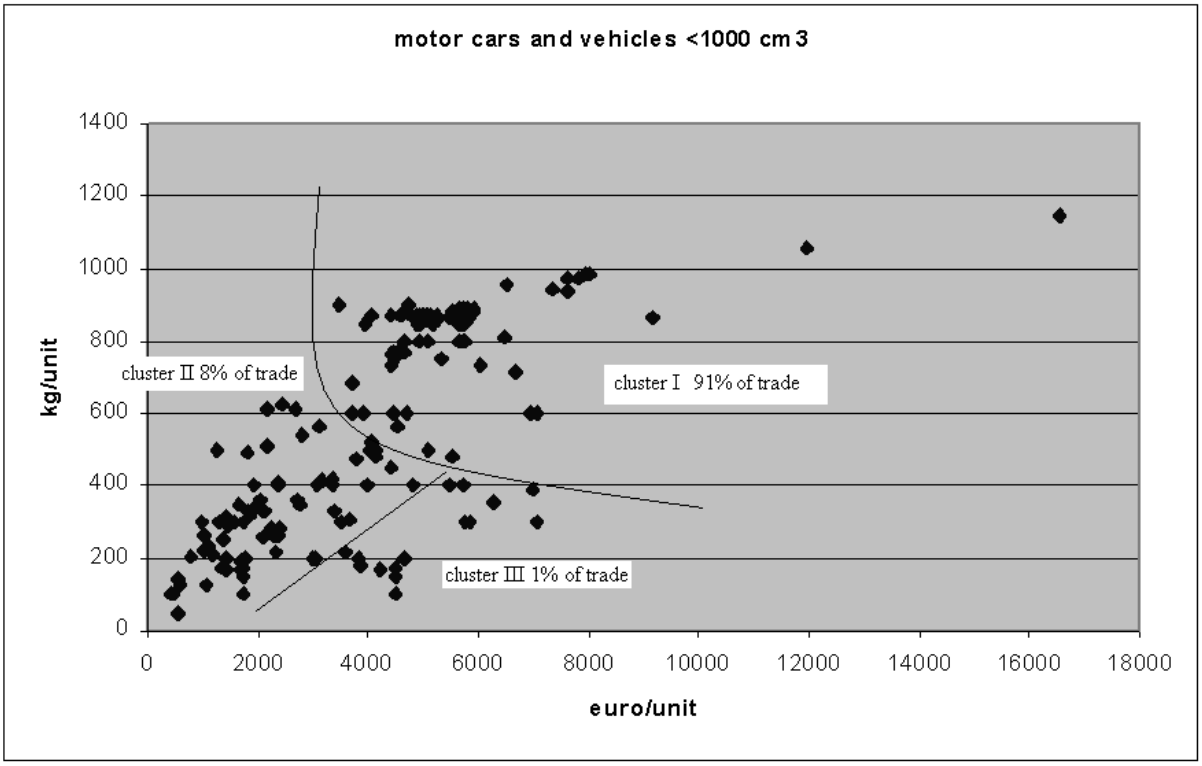
- Definition of validation intervals. In case that the main cluster represents more than 90% of the trade value, then we obtain one interval with the information provided in this main cluster. Two methods can be applied: Traditional based on average and standard deviation or robust based on median and MAD or quartiles.
- In case the main cluster represents less than 90% we define a separate interval for the second size cluster (eventually for the third etc.). If the clusters intervals overlap, then they are replaced with one which is the union of both.
- After some investigations, the clusters defined above can be considered as different products and used for the elaboration of indices without further cleaning or to assist classification works (evolution) or to provide keys for backward or forward estimations related to product evolution.
- After some investigations, the clusters defined above can be considered as different products and used for the elaboration of indices without further cleaning or to assist classification works (evolution) or to provide keys for backward or forward estimations related to product evolution.



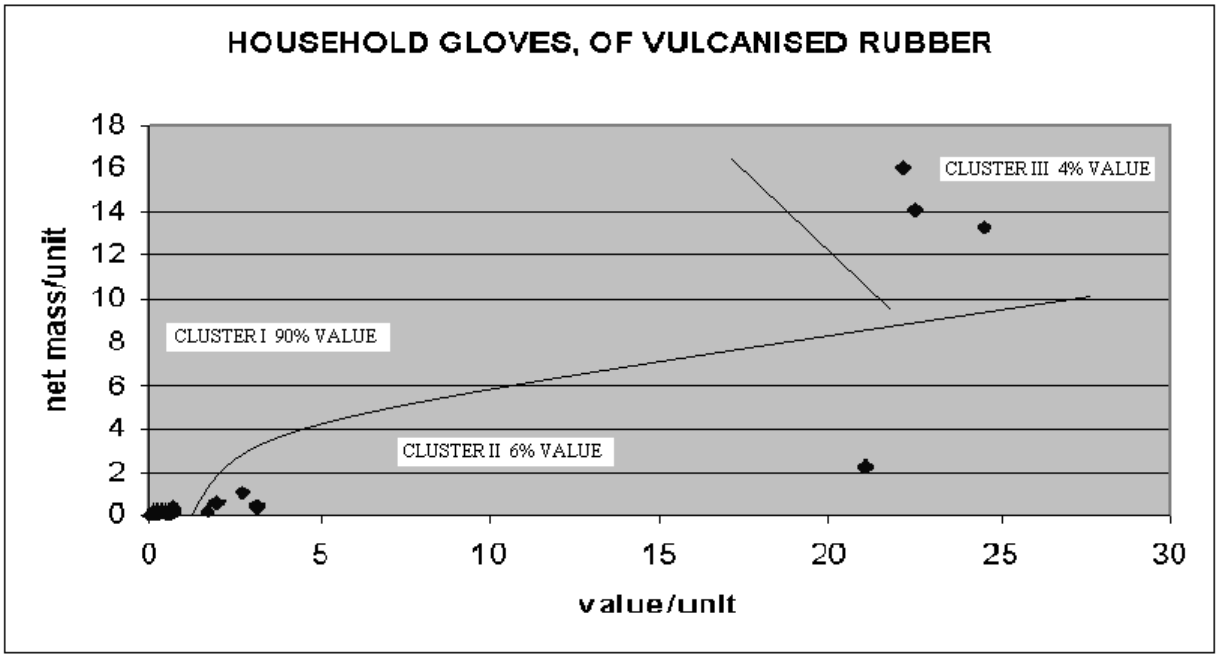
26. Some examples of clustering method:

Example of two or three clusters without outliers

2 clusters R-square 46%
3 clusters R-square 69%



Example of two or three clusters with outliers
2 clusters R-square 60%
3 clusters R-square 86%


HOUSEHOLD GLOVES, OF VULCANISED RUBBER
net mass/unit
value/unit
CLUSTER I 90% VALUE
CLUSTER II 6% VALUE
CLUSTER III 4% VALUE
0
2
4
6
8
10
12
14
16
18
0
5
10
15
20
25
30