

Mirror Outlier Detection in Foreign Trade data

Spyros Liapis¹, Michael Ashbrook², Markos Fragkakis³, George Petrakos⁴

¹Agilis SA, e-mail: Spyros.Liapis@agilis-sa.gr

²EUROSTAT/UNIT-G3, e-mail: Michael.Ashbrook@ec.europa.eu

³Agilis SA, e-mail: Markos.Fragkakis@agilis-sa.gr

⁴Agilis SA, e-mail: George.Petrakos@agilis-sa.gr

Abstract

This paper explores Mirror Outlier Detection in Foreign Trade data. More specifically it presents the MOD application developed in Eurostat/Unit-G3 that performs detection on data extracted from COMEXT. The Median Absolute Deviation method is used for outlier detection. It illustrates how the outliers are characterised based on the mirror series. It outlines how MOD handles sort series due to product changes. Also it presents the methodology for classifying outlier due to errors in the series dimensions and not in the observation values. Outcome of MOD is presented in graphical form. Technical information on the design and implementation of the application are given. A new platform for visualising and analysing outlier data, stored in MOD's repository, is proposed. The visualisation is based on a multidimensional viewer supporting OLAP functions. Data analysis is based on log-linear models for extracting inferences from the data.

Keywords: outlier detection, data visualisation, data mining

1. Introduction

Eurostat collects data from the member states for Foreign Trade on a monthly basis. The volume of this data is huge since the products codes declared by each member state are above 10 thousand. This necessitates the efficient storage and validation for further exploitation and dissemination of the data. Consumers of the data are the European Commission, national authorities, international authorities, private companies and other. For this reasons the quality of the data is crucial.

Quality in Foreign Trade statistics within the European Statistical System is based on the following dimensions (Eurostat-F2, 2004):

- Timeliness deals with the publication calendar, reference period, etc.
- Accuracy deals with the closeness between the value finally retained and the true unknown population value, e.g. exclusions, thresholds, outliers, non-response, adjustments, controls and corrections, confidentiality, etc.
- Accessibility deals with availability, ease of access to data, different formats and conditions of data distribution, etc.
- Clarity deals with ensuring data is adequately documented, assistance in using and interpreting the data, etc.
- Comparability deals with conceptual differences between sets of trade statistics over space and over time.

- Coherence deals with to what extent statistics originating from other sources (such as balance of payments, national accounts, etc.) are compatible with trade statistics.

In this work the use of outlier detection, which falls into the accuracy dimension, is employed as a mean of data quality improvement. Foreign Trade data are in the form of monthly time-series, which means that outlier detection consists of the identification of observations with exceptionally diverging values (either larger or smaller) from the rest of the time series.

There are several methods for detecting outliers that share the common approach of specifying threshold values, which can distinguish between a normal and an outlier value. These methods could be classified into the following broad categories:

- Specification of fixed thresholds estimated from experience gained from historical data.
- Descriptive analysis: specification of thresholds by using descriptive statistics, e.g. mean value and standard deviation or median and quartiles of the inspected time series, the MAD method, etc.
- Model based methods: specification of thresholds using models to estimate means, standard deviations, etc. The proper model to fit the available data may be selected from a variety of alternatives like non-parametric regression models (e.g. the application of THEIL regression (Sprenst, 1993)), simple linear model (OLS method) or its extensions (WLS, log-linear regression, etc.) and time series models (e.g. ARIMA).

The method preferred in this work is the MAD (Median Absolute Deviation) due to its robustness that has been proved from its use in an existing application for outlier detection in Eurostat. In this paper is presented the successor of this application. The new application is called MOD (Mirror Outlier Detection) that improves on the old application's features.

The predecessor of MOD performs outlier detection using MAD in single time series, which resulted in poor characterisation of outliers (no indication of whether the outlier could be due to real fluctuation in the data, or due to error in a specific reporter observation). Moreover the old application could not handle discontinuities in series due to product changes. The detected outliers were not examined on whether they were caused by errors in the dimensions of the time series or in the observed values. Also another problem was that the reporting was poor and not easy to read. Another problem was that there was no common repository of outlier data, which would permit data exploration and analysis overtime. Finally the old application was developed on a proprietary platform (SAS) that limited its portability, maintainability and performance.

In the following sections of this paper the MOD application is presented. These highlight how MOD deals the shortcomings of the previous application and the new features added, by exploiting the mirror trade flows (i.e. the series declared by the partner county), as indicated by the application's name.

The presented work is organized as follows. Section 2 presents the methodology adopted by MOD. Section 3 analyses the application in terms of software design and

implementation and illustrates the outcome of the application in graphical form. Section 4 proposes a new platform based on the MOD application, which will enable outlier data visualisation, exploration and analysis. The visualisation and exploration is based on an interactive OLAP tool. Moreover it proposes a statistical methodology is proposed for the extraction of useful inferences from the detected outlier data. Sections 5 and 6 present the conclusions and proposed future work.

2. Methodology

Foreign Trade data consists of time series, which are collected on monthly basis, and whose each observation concerns a month period. For each month, 3 measures are observed: Value (in 1000€), Quantity (in 100Kgs) and Supplementary Quantity (measured in a supplementary unit which depends on the product i.e. pairs for shoes). The dimensions that uniquely identify an observation – along with the reference period and observation variable - are the Reporter, the Partner, the Product and the Flow (import or export).

Univariate outlier detection is performed on the series of each observation. The method used by MOD for the detection of outliers is the Median Absolute Deviation (MAD) (Tukey, 1993). The MAD is a measure of statistical dispersion, more robust than the sample variance or standard deviation in the presence of outliers; it also exists for some distributions, which may not have a mean or variance (e.g. Cauchy distribution). Furthermore, since it is nonparametric, no assumption on the distribution of the data (e.g. normality) is needed.

To determine whether an observation x_i is an outlier or not MOD uses the T statistics:

$$T_i = \frac{|x_i - M_1|}{M_2} = \frac{|x_i - M_1|}{\text{Median}(|x_j - M_1|)}, \quad j=1,2,\dots,i,\dots,n$$

where, M_1 is the median of the observations and M_2 is the MAD of the data set (including the inspected observation x_i). The median absolute deviation (MAD) is actually the median of the absolute deviations of the observations from the median of the data set. The observation x_i is consider to be an outlier if T_i is greater than 5.

The predecessor of MOD also used MAD for the detection of outliers, however, it did not take into consideration the Foreign Trade mirror flows for their characterisation. The mirror flow of a time series is the same time series, as reported by the partner country (in practice, the mirror of a series is a series where the flow is opposite and the reporter and partner countries are reversed). The characterisation of an outlier provides critical information on its nature, most importantly whether it is due to an actual fluctuation in the product trade, or due to an error by one of the data providers. Having detected the outliers in a time series, each outlier can be characterised based on the mirror series:

- If the mirror series does not exist (is not in the dataset) and the mirror flow is declared to be confidential, the outlier is characterised as “black”.
- If the mirror series does not exist and is declared to be confidential, the outlier is characterised as “pink”. This is a normal side effect of data confidentiality.

- If the mirror series is in the dataset, but there is no mirror outlier at the same reference period, the detected outlier is characterised as “red”. This means that the outlier is probably caused due to an error in the reported data.
- If a mirror outlier is also detected in the mirror series with deviation of the same sign (i.e. both outliers have either larger or smaller values than the median of their series), then both outliers are characterised as “green”. This means that the outliers are possibly due to a valid fluctuation in the given product trade.
- If a mirror outlier is detected, but the two deviations have opposite signs, both outliers are characterised as “violet”. This means that the reported data are invalid.

There are a number of cases where the nature of the processed data makes the detection of outliers difficult. One such case is the existence of discontinuities in the time series due to changes in the codes of products. These changes take place because the product Combined Nomenclature is updated annually, resulting in single product codes splitting to new codes, multiple codes merging in a new one, or single codes replaced by new ones. There are also cases where the mapping of old codes to new codes may be “many to many” relation. The described product code changes result in the segmentation of what would be a single time series for a product or a group of products to multiple time series with fewer observations. The application of MAD, however, relies on the existence of a certain number of observations for the algorithm to run reliably. MOD tackles the problem by using product code changes metadata, which are extracted from COMEXT, to “compose” extended time series of two or more product codes. These extended series consist of the concatenation and aggregation of all the series that belong to the same product-group according to the code mapping. The extended time series are identified with a product code that has a “_MERGED” suffix and are separately included in the produced reports.

MOD is also able to handle certain cases where the error does not lie in the variables of the series (Value, Quantity, Supplementary-Quantity), but rather in the dimension values e.g. an observation value was attributed to a wrong country or product code. There are three types of errors that may be checked: copy or swapping of values between series and hidden green outliers. A strong indication that such error has occurred is the existence of outliers that are red in all three variables, which narrows significantly the number of series to be checked. In order to indicate whether a red outlier is due to swapping, MOD finds all the other red outliers at the same reference period, and consecutively exchanges the outlier with the others. If both outliers disappear (detection is re-applied) it is considered an indication that the two observations may have been attributed to a wrong series. In order to check for observation copying, a simple similarity check is performed for all 3 variables between the outlier and observations of similar series (similar stands for series with only one dimension different, i.e. product code). The hidden green outlier refers to the case where a red outlier occurs because the respective value in the mirror series is declared earlier or later (depending on the flow). This is common when there are delays in declaring a flow for specific products (e.g. due to transportation), in which case observations in both a series and its mirror will be detected as red outliers. In order to detect such cases, the mirror flow of a series with a red outlier is checked for a red outlier (in all 3 variables) for a period of 1-2 months before or after, depending on the flow of the series.

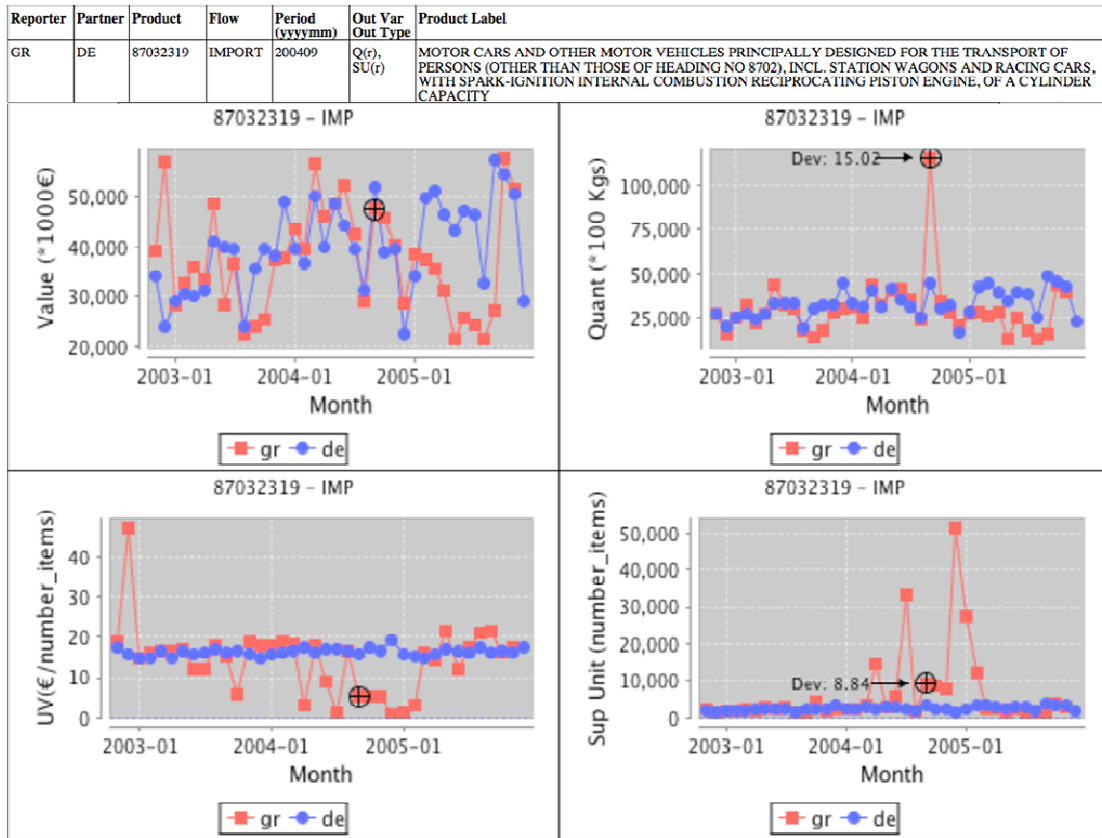
3. Application

The implementation of MOD is significantly different from the way its predecessor was implemented, which results in a number of advantages. Firstly, MOD is written in Java, which makes it independent of platform, in contrast to its predecessor, which was developed in SAS. Furthermore, MOD has improved on the performance of the detection, since the full mirror detection (50 million series or about 2GB of data) can complete in one day while the old application required the same time for single detection for just the data of Germany. Finally and most importantly, instead of storing the outlier data in text files, MOD stores the outlier data into a relational database. This is probably the most significant difference, as it allows further processing and reporting at any time, for any detection execution.

From an architectural perspective, the design of MOD is layered. The top layer is a CLI (Command Line Interface), through which the implemented functionalities are made available to the operator. Underneath lies the layer that contains the application logic, where all the functionalities are implemented (detection, characterisation of outliers, reporting). Finally, there is the data layer, whose role is twofold: it is responsible for the reading the application parameters, metadata and data files, and also for reading and writing outlier data in the MOD database (either Oracle or MySQL) over the JDBC protocol. Such data are the outlier values and other metadata, along with the series that contain them and their mirror series.

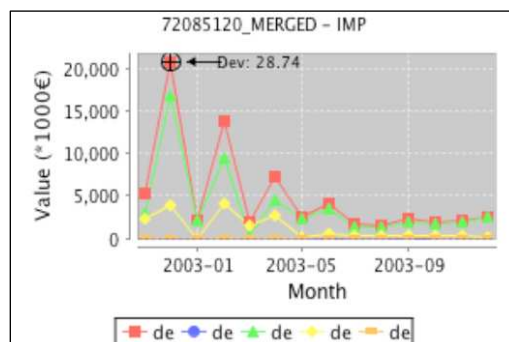
The input used by MOD is a number CSV (comma-separated values) text files that are all extracted from COMEXT. They contain several types of data, like nomenclatures, data update dates, product code changes, confidentiality information and most importantly, the actual trade data. The use of text files was adopted because at the time of the development COMEXT did not have a stable interface for querying. Furthermore, the CSV format is a convenient format to generate from other data sources besides COMEXT, which can facilitate the extension of MOD to other domains.

After the detection and classification functionalities are executed, MOD produces two types of reports: detailed and summary reports in PDF and CSV formats. A detailed report is generated for each reporter, which contains a list of the most important outliers ordered by estimated significance. For each of the reported outliers, there are detailed graphs showing its position in the time series, its relative deviation, and its mirror series. This type of report is sent automatically to country correspondents by e-mail and can be used to indicate and fix possible errors in the data.

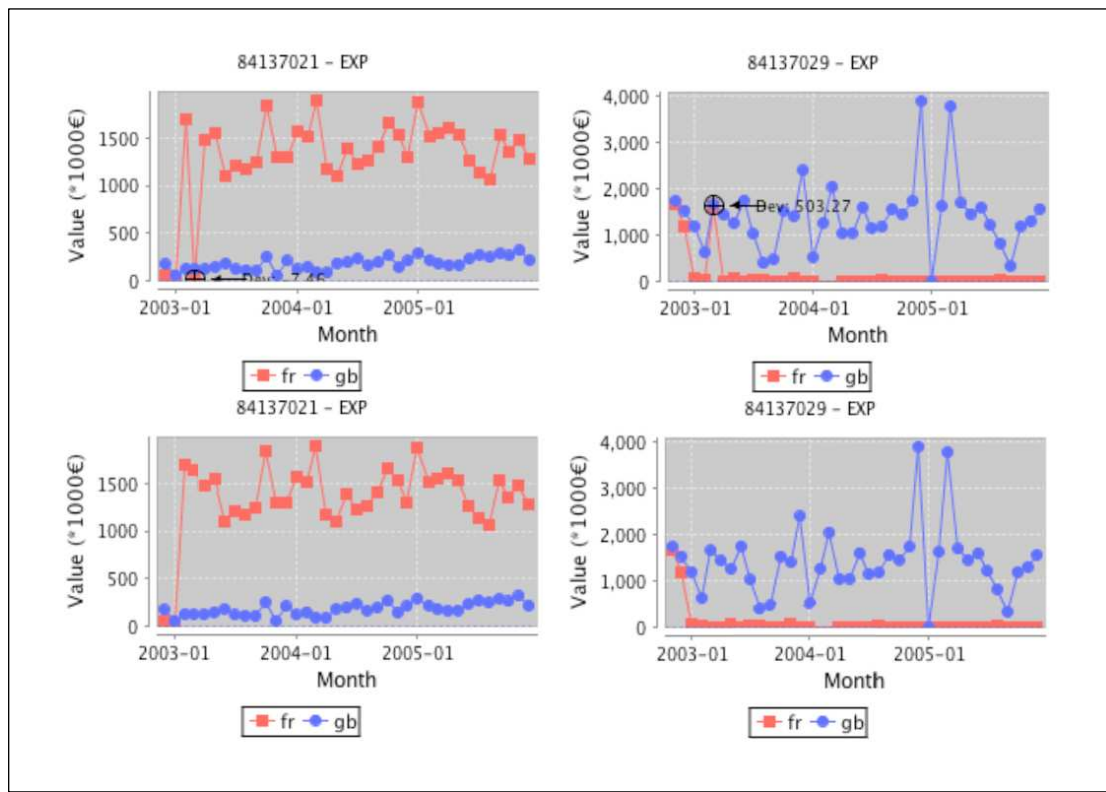


The above figure shows the details of a detected outlier. The details on the top of the figure show details, like the reporting and partner country, the product code and the reference period when the outlier occurred. The reader can also see the variables where the observation is an outlier (here Q for quantity and SU for supplementary unit quantity), as well as the colour of the outlier for each variable (here red for both). The four diagrams show the position of the observation in the time series, as well as the deviation from the median, where the variable is an outlier.

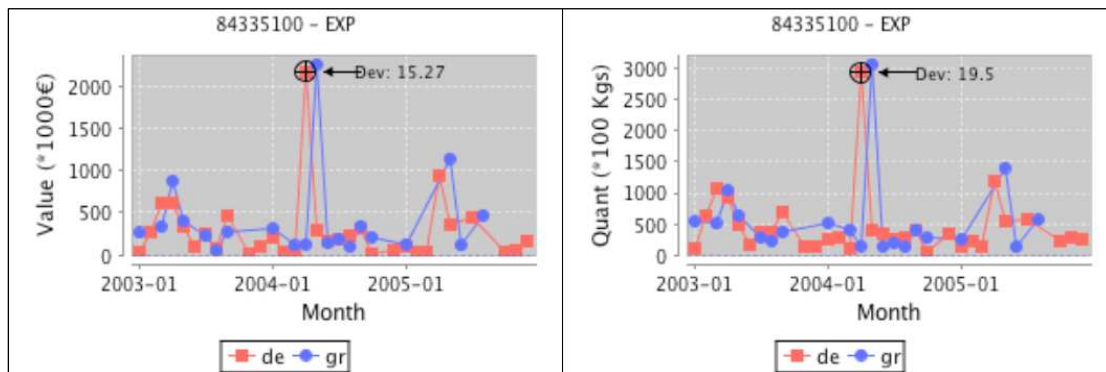
Outliers of **extended (merged)** time series are also reported in a separate section. The produced figures are the same as the ones for normal outliers, with the exception that here the product code has the suffix “_MERGED” and the description field contains the list of the product codes whose time series have been merged to produce the extended time series.



The figure above shows the merging of several time series (green, yellow, orange) into a single one (red). This aggregation takes place for all variables (value, quantity, supplementary unit quantity).



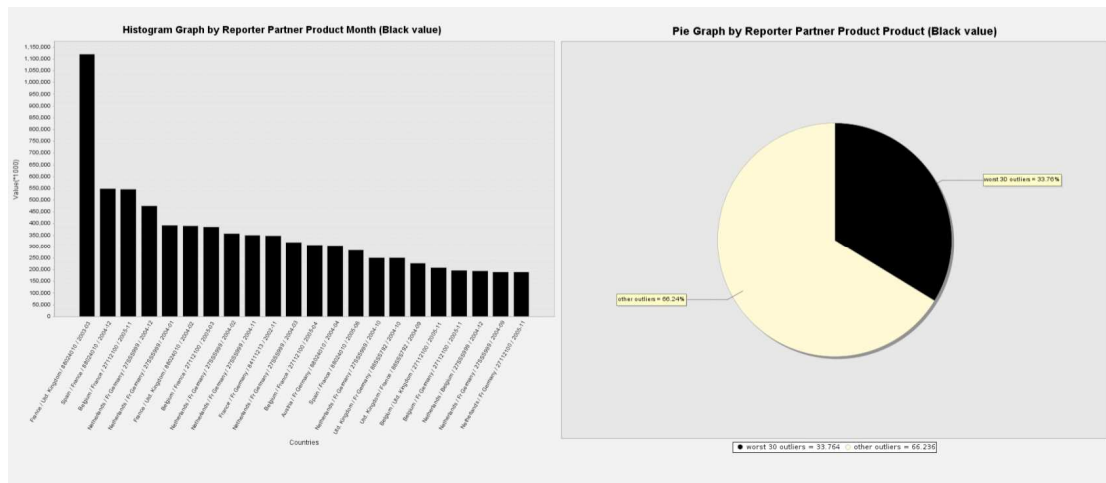
The above figure shows the discovery of a possible swapping between two outliers in the data. The charts in the top row show two series containing with marked red outliers in the Value variable. The bottom charts show the two series after the swapping proposed by MOD takes place.



The above figure shows the characterisation of detected red outliers as possible hidden green outlier due to time delays.

The summary report consists of tables that aggregate the detected outlier values for specific combinations of reporter, partner, period and product. This report is very useful for identifying specific subsets of data that appear to have the poorest quality. For instance, if a specific product appears to have the largest aggregated values of red

outliers, it is apparent that further investigation and reporting on this product would improve data quality significantly.



The above figure shows the types of diagrams that are produced for each combination of reporter, partner, period and product. In the specific case, both diagrams show the “worst” (largest) 30 aggregates of the variable “Value” from all combinations of reporter, partner, product code and reference periods, where the outliers are black. The pie chart on the right shows the percentage of these 30 reported aggregates from all the aggregates of the same type.

4. New outlier visualization and analysis platform proposal

MOD detection runs on a monthly basis on the latest reported data in COMEXT, thus every month hundreds of thousands outliers are detected and stored in MOD’s database. This huge volume of outliers is not exploited as much as possible since the static country reports provide useful information but just for a small set of the highest valued outliers. Even summary reports that provide aggregated information in various break downs is not so handy as the interactive exploration of data. Also data are not analysed to provide further inferences on the outliers.

Therefore there is a need for a new platform that will enable the interactive data visualisation and analysis. In this paper we propose a platform that will consist of two major modules the data visualisation and the data analysis. Data visualisation will be responsible for the interactive exploration and visualisation of data. On the other hand the Data analysis module will be responsible for providing the statistical methods for data analysis (e.g. data mining).

As regards the visualisation, since the outlier data are multidimensional, a multidimensional data viewer offering OLAP (On Line Analytical Processing) functions would allow for interactive exploration of data. This viewer will present the data in 2-D tabular form that will offer the possibility to dynamically place data dimensions in the axis of the tabular form or leaving them out of it. Practically this realises the slice and dice OLAP operations. When more than one dimensions are placed in one axis automatically the dimensions are multiplexed. The viewer will offer sorting of data in the grid and create charts like time series, bar charts and pie

charts. Furthermore this viewer will enable the roll-up and drill-down e.g. for the product it would aggregate the view to codes in higher level or then drill them down. This requires the dimension to be related with a hierarchical nomenclature (e.g. CN8 for the products). If the classification is not hierarchical then only could be rolled up to the total value by practically aggregated it out of the view. Besides the default sum operation it could offer max, min, average, mean, median, count operations.

Moreover it will support estimated variables from the existing variables of the data, thus it will possible to produce quality indices over time. Thus, besides visual exploration of data, this platform will be possible to provide summary statistics of the outlier data.

Such a visualisation tool has been developed in Java in other projects carried out by the authors but it needs to be customised for the use in the proposed platform.

The visual data exploration would easily pinpoint data with poor quality and possible reasons for this behaviour (e.g. specific products involved in fraud like cigarettes, pharmaceutical preparations, and electronic devices). Data exploration will guide the user to subsets of data, which could be further processed with data mining methods to find hidden relations between outlier data.

A class of log linear models are proposed here, having as explanatory variables the reporter, the partner, the product and the outlier type in different levels of aggregation/classification. In these models the response variable will be the number of outliers, assumed to follow Poisson distribution and the natural logarithm of these cell counts is expressed as a linear combination of the effects of exploratory variables. Specifically the model is the sum of a constant term, the main effect variables and the up to k^{th} order interactions. Applying this type of model to MOD data, we aim to discover the structural relation of these exploratory variables and their interactions with the occurrence of outliers in our time series. Modelling one product at a time we define a four-dimension model with the first order interactions:

$$m_{(ijkl)} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{4(l)} + \mu_{12(ij)} + \dots + \mu_{34(kl)}$$

Using ML estimates of cells we can identify “combinations of variable categories that shows large deviance from a hierarchical model apply to the whole array” (Bishop, 1975), by examining the goodness of fit of each cell. Using these analyses over time we can also assess the quality of data and identify trends in outlier appearance.

5. Conclusions and future work

MOD enables the detection of outliers in Foreign Trade data extracted from COMEXT. It characterises outliers according to the mirror series. Further it handles short series due to product code changes and checks for outliers caused from errors in the series dimensions. It produces report per country with the greatest outliers and summary report for all data. A platform offering visualisation and analysis tools for utilising outlier data was proposed, that is it to be implemented in the future.

Finally as future work for the MOD platform it would very interesting to extend MOD to international data (e.g. foreign trade of China, USA, Eastern Europe) in order to compare trade data in international level since now it performs mirror detection into EU internal trade data. This will possibly need MOD to adapt to different nomenclature systems (e.g. HS product nomenclature) and reporting periods (e.g. quarterly data). Besides Foreign Trade data MOD can be used in other type of data where mirror flows of data exists e.g. balance of payments, migration flows.

References

- Hoaglin, Mosteller, Tukey (1983) Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, Inc., Chapter 7. Pages 211 – 246, Chapter 9. Page 291.
- Sprent, P. (1993) Applied Non-parametric Statistical Methods, Chapman & Hall, Chapter 8, Pages 188 – 208.
- Eurostat F2 – International Trade (2004), Quality Report on International Trade Statistics
- Chambers R., Hentges A. & Zhao X. (2004) Robust automatic methods for outlier and error detection, Journal of the Royal Statistical Society 167, 323-339
- Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975) Discrete Multivariate Analysis: Theory and Practice, MIT Press, Massachusetts and London
- Eurostat, COMEXT database of external trade statistics