

Managing of huge administrative data: the example of EU external trade statistics

Berthomieu Anne

EUROSTAT, Bâtiment Jean Monnet L.2920 Luxembourg,
anne.berthomieu-cristallo@ec.europa.eu

Abstract: The EU external trade statistics represent a huge volume of data collected each month through Customs formalities. Such a volume impacts each stages of the statistical process, i.e. the collection, the validation and the dissemination of data. Only essential information should be collected and processed at the most detailed level. Dedicated tools are needed in order to control the data quality and raise errors. Wherever possible, erroneous information should be automatically corrected but the human intervention remains absolutely necessary in case of key issues. What is a key issue should be defined according to the variable and the stage of the statistical process. The dissemination should concentrate on key information answering the needs of the different categories of users and optimise the data consultation.

Keyword: Managing huge data

1. Introduction

The aim of this paper is to show to which extent the data volume impacts the processing of the statistics on the trade in goods of the European Union (EU) with third countries – so-called extra-EU trade statistics.

A brief description of the extra-EU trade statistics: For each Member State and trade flows – imports and exports –, each movement of goods through the EU frontiers is registered in terms of statistical value (value in euros of the goods when crossing the EU frontier, i.e. including the transport and insurance costs up to the EU frontier) and quantity (kilograms but as well other units when relevant like for instance the litter or the piece). The goods should be identified through a product nomenclature holding about 19000 10-digit codes for the imports and 11000 8-digit codes for the exports. The country of final destination in case of export or of origin in case of imports should be reported with as well the nature of the transaction (transfer or not of ownership, trade for processing like packaging or labelling, etc) and the possible preference in case of imports (less Customs duties for specific goods originating from specific trading partners). In addition, information should be given on the transport means (in particular, the mode of transport: by road, sea, air, etc).

Who does what in the statistical process? The EU Member States collect the extra-EU trade data through Customs procedures and carry out quality controls before providing Eurostat with data compiled according to the EU concepts and definitions. Eurostat controls the data completeness, validity and consistency before making data available to

the users through its various dissemination channels (paper and electronic publications, databases, DVDs, etc).

Some volume indicators: According to the Member State and the trade flow, up to 3.5 millions of Customs lines are processed each month and the average length of one record is between 110 and 1300 bytes. Regarding the data transmission to Eurostat, around 1500 files were sent in order to report the 2008 provisional and final monthly figures. For the whole year 2008, Eurostat processed 3 millions of records, which represents an annual data volume of 4 gigabytes.

2. Data collection

The data sources: The Customs Authorities of the Member States use the Single Administrative Document (SAD) as the documentary basis for customs declarations on the trade in goods with non-EU countries. Up to 38 fields of the SAD - over a total of 54 - are further used by the Customs and/or by the National Statistical Institute in order to compile extra-EU trade statistics. With the aim to reduce the operators' burden as much as possible and increase the efficiency of the procedures, the SAD can be replaced by several simplified declarations gathering only some key information and followed, in a later stage, by an exhaustive recapitulative declaration. Over the EU and for the year 2007, about 70 millions of SAD were collected in order to cover the trade relating to 110 millions of items. Around 150 millions of supplementary items were reported through simplified declarations.

The promotion of electronic formats: Declarations to Customs can be made on paper or in electronic formats. Considering the data volume, the electronic formats are systematically promoted. In particular, from July 2009 onwards, the declarations on exports of goods should be made available only via electronic means, either using Customs web tools or by data exchanges through networks.

A simplification wherever possible: The EU legislation has made optional the collection of some statistical elements – like, for instance, the region. Therefore, it is up to the Member State to decide to which extent it is worth collecting and compiling this type of information. In addition, the Member States are free to apply a statistical threshold in order to avoid processing all the Customs declarations at the most detailed level. This statistical threshold shall be fixed, for each type of goods, so that imports or exports exceeding EUR 1000 in value or 1000 kg in net mass are collected for the production of trade statistics. In other words, only the goods movements above a certain value should be covered by the extra-EU trade statistics at the most detailed level, i.e. with the product identification, the mode of transport, the nature of the transaction, etc. For the other movements, the EU legislation requires only a global estimate.

3. Data validation

Who should do what? To be optimal, quality controls should be implemented as early as possible in the production line starting from the data collection within the Member

States and ending by the data dissemination in Eurostat. According to the stage of the process, the validation rules may address different issues or lead to different corrections: (1) in Customs local offices when the SAD is made available, (2) in the Competent National Administration (CNA) when data are aggregated and compiled for national and Community purposes and (3) in Eurostat before the data dissemination through EU channels. Eurostat and the Member States are working hand in hand toward the implementation of a harmonised approach in order to avoid duplicated validations or performed too late in the process.

A need for automatic validation and correction procedures: Because of their volume, the extra-EU trade statistics can only be validated through automatic procedures. In Eurostat, all the data files transmitted by the EU Member States are automatically processed by a dedicated production system. This system points out missing, invalid and inconsistent entries for all provided statistical variables and detects outliers in trade indicators – statistical value and net mass. Some errors are considered as "fatal", which means that the automatic data processing stops and a human intervention is then necessary for investigation and decision.

Fatal errors are mainly referring to missing, invalid or inconsistent entries impacting key variables, i.e. variables considered as essential for the users and/or on the basis of which other variables are validated (examples of key variables: the type of trade flow - import or export -, or the type of data - collected or estimated by the NCA -). Outliers are only considered as fatal errors when they are observed at total level – i.e. total export and import values. They are detected on the basis of pre-defined intervals derived from simple linear regressions: $\pm 30\%$ of the prediction for the statistical value and $\pm 50\%$ of the prediction for the net mass. In most cases of fatal errors, a data revision is requested from the Member State concerned. Only in exceptional cases, the erroneous entry is manually corrected or the interval manually extended.

Non fatal errors correspond to missing, invalid or inconsistent entries impacting non key variables. Such errors are automatically corrected using default values and listed in a report indicating the original and corrective values. Outliers observed in the trade indicators aggregated at the 2-digit level of the product nomenclature are as well listed in a specific report but not automatically corrected. A "human" assessment of both reports is necessary before going up to the dissemination to users.

Further investigations on automatic correction methods:

For the time being, Eurostat applies only simple rules for automatic correction based on default values. Regarding the trade indicators (statistical value, net mass and supplementary quantity), the automatic correction simply consists in replacing any negative values by zero. Nevertheless, different advanced statistical methods have been tested in order to assess to which extent the indicators might be automatically imputed in case of outliers. The general conclusion was that Theil's regression performed better and with greater consistency than all the other methods considered: ordinary linear regression, log-linear regression, ARIMA models for time series and RegARIMA. An additional conclusion is that it is worthwhile to apply Theil's regression using another indicator as predictor – like, for instance, the net mass as predictor for the statistical value- instead of using time. This is the consequence of the lack of stability of the time series due to the high level of details: in particular, records are impacted by the annual evolution of the product nomenclature and moving trade patterns.

4. Data dissemination

Selection of key indicators: The data volume is not an IT issue for the dissemination of the extra-EU trade statistics although it is an important element to be considered when taking decisions in terms of data storage and management. Rather than the data volume, it is more the level of details which should be taken into account in order to optimise the data dissemination to users. In Eurostat's publication databases, key information are shown through different views of a single dataset. In each view, maximum 6 statistical dimensions are proposed with the aim to avoid the extraction of too many empty cells. Each view focuses on specific elements of the detailed extra-EU trade statistics identified as of high interest for certain categories of users.

Further compilations: For some specific purposes – in particular, economic analyses -, supplementary indicators need to be compiled: in particular, the seasonally adjusted trade values and the unit-values indices. Regarding the seasonal adjustments: the seasonal adjustment process needs to concentrate on key time series in order to remain manageable on a monthly basis. A way to further reduce the number of series is to use more aggregated product nomenclatures which are, in addition, more adapted to economic analyses - for instance, the Broad Economic Categories (BEC). Regarding the indices: Their compilation is only possible thanks to the details and the high degree of completeness of the extra-EU trade statistics. Nevertheless, their dissemination should follow the same principle than the one applied for the seasonally adjusted data: selection of key time series and use of product nomenclatures answering better the users' needs.

Data consultation: Over the year 2008, more than one million of data extractions have been carried by nearly 2500 registered users. The number of extracted lines was about 270 billions. The performance of the data consultation is ensured by a strict management of the number of extractions running at the same time – maximum 50 – and as well by a limitation of the extraction size: maximum 5000 items can be selected in each statistical dimension; maximum 240000 cells can be extracted in interactive mode and 1.2 million in batch mode.