Database Ulisse: Methodological note

September 2018

Table of Contents

Presentation

This document describes the methodological choices that have been made to build the database Ulisse. The data of exports and imports declared by the various countries of the world are a rich database of information content. Using these data it is possible to have information on the dynamics of different markets at a very high level of detail . And it is also possible to know at what price a given good is sold on a specific market and, above all, if some operators on the market pay or not a *premium price* for higher quality offered. Finally, it allows to assess the competitiveness of different countries, highlighting their strengths and weaknesses. Often, however , this information is not immediately accessible from a simple reading of the data. The statements of different countries sometimes contain data that does not correspond to the actual flows of foreign trade. The main reasons are as follows:

Estimation procedures : not always data are collected when the goods arrive at the different customs borders. It is the case, for example, of European Union countries' foreign trade flows, whose values are estimated by different national statistical institutes on the basis of VAT returns made by the various companies. Generally, however, these estimation procedures are of high reliability.

Product classification : not all countries use the same product classifications and, therefore, flows relating to similar products can be attributed to different product codes. In recent years, however , there was a strong process of harmonization of different classifications, mainly thanks to the work done by the World Customs Organization (www.wcoomd.org) that has developed a *Harmonized System* classification, which is used for almost 150 countries.

Confidentiality : sometimes foreign trade flows by country and product can be secreted because they are considered confidential. Each country follows its own rules, with a higher incidence of secreted flows for smaller countries. Among the countries of the European Union some ones (i.e. Denmark, Finland, Austria) present a ratio of secreted flows more than 5% of their total exports.

Delays : some countries in the developing world, with a less well-established administrative structure, tend to produce statistics relating to their foreign trade flows with delay, in some cases consisting in even a few years. The UN is committed to support these countries by financing projects that aim to improve the procedures for collecting, processing and publishing foreign trade data.

In order to extract meaningful and reliable information from the data base, it is therefore necessary to use a variety of methodological tools designed to distinguish the *information content* of the phenomenon from the *statistical noise*.
It is therefore essential to build analytical procedures that, on the basis of what has just been described, best use the many advantages inherent in the international trade data, namely:

High Numerousness : most countries publish monthly statistics of foreign trade by product code and by partner country. It follows that, each month, each country publishes data related to several million trade flows data: the big

number of data makes it possible, therefore, to easily identify possible outliers.

Data refer to the population : foreign trade statistics affect almost all trade flows. This allows to avoid all the problems regarding sampling and measurement errors that can arise from their use.

Double Statements : each foreign trade flow is declared twice: once by the exporting country and another time by the importing country. The double declaration, if properly used, allows to increase significantly the reliability of the measures relating to various phenomena.

These features concerning foreign trade data allow to make effective use of some methodologies of *data mining* able to isolate satisfactorily the *information content* of the various phenomena from their *statistical noise*. This document describes the different methodologies used in the construction of the Data Warehouse Ulisse. These methodologies include:

Outliers Detection : methods aimed at identifying data affected by significant measurement errors and their replacement with values more consistent with other observations;

Missing Data Detection : methods that identify trade flows for which some measures are missing. These methods allow to estimate for the missing data a value consistent with the information set.

Normalization : methods that allow to compare the different sizes of the same phenomenon (i.e. flow declared both by the country of origin and by the country of destination), producing for each variable a unique measurement.

Conjuncture : methods capable of producing a preestimate of an annual figure, using the best conjuncture information available.

Price Ranges for differentiable goods : methods that allow to assess whether a certain good may be subject of quality differentiation. In the case of a differentiable good, the procedures developed by StudiaBo give a specific range of price (quality) for each flow, taking into consideration the comparison between its price and that of the other flows for the same good.

Quantity at constant prices : foreign trade flows declarations relate both to values and quantities. The ratio between the two variables consists of the so-called *average unit value*, often identified with the price. These values also incorporate qualitative changes affecting the product. In order to detect these changes, StudiaBo built the variable *Quantity at constant prices* (Q), expected to include also quality changes in the product. The prices derived from that, given the relationship between values and quantities at constant prices, can therefore be considered more reliable than the average unit values, because they tend not to be affected by quality changes in the product.

We believe that the methods used by StudiaBo and hereafter described allow to produce an highly informative dataset.

The sources of database Ulisse

UN Comtrade

The database Ulisse was constructed from available information from different sources of economic analysis, first of all the United Nations Statistics Division. This entity has built and has been maintaining the Comtrade database (http://comtrade.un.org/db/). UN Comtrade brings together the annual foreign trade flows of more than 170 countries, detailed by code-level product. The Statistics Division of the United Nations has estimated that UN Comtrade covers more than 95% of world trade. For each flow of foreign trade (as a declaration of a country to its commercial partner in a given year), UN Comtrade reports the dollar value and the quantity exchanged (which is expressed either in kg and / or a supplementary unit, different from product to product).
Harmonised System (HS)[1] is the product classification used in the database Ulisse.

Other sources

The objective of having a database capable of supporting the analysis also of more recent events has necessitated the development of a procedure for the periodic updating of the database Ulisse. This result was made possible by making use of economic information made available from the following other sources:

- **Eurostat Comext database**: monthly foreign trade flows declared by the EU and EFTA countries (http://epp.eurostat.ec.europa.eu/);
- **UN Monthly Comtrade database**: monthly foreign trade flows declared by UN countries partecipating to the UN Monthly Comtrade project (http://comtrade.un.org/monthly/);
- **UN Monthly Comtrade database**: monthly foreign trade flows declared by UN countries partecipating to the UN Monthly Comtrade project (http://comtrade.un.org/monthly/);
- **U.S. Census Bureau - db UsaTrade**: monthly foreign trade flows of U.S. companies (https://usatrade.census.gov/).

Eurostat - db Comext

Eurostat Comext (http://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?sort=1&dir=comext) is the database containing the monthly statements of foreign trade reported by EU and EFTA countries.
For each monthly flow of foreign trade, db Comext reports the measures both in monetary values and in volumes (kilograms and / or other supplementary unit measure). Data are updated monthly at fixed intervals, with a delay of about six weeks for flows reported to partner countries outside Europe and about 10 weeks for flows reported to European countries.
EU member states are using a very detailed level of customs classification, called the Combined Nomenclature, at 8-digit level (CN8). CN8 consists of a further specification with respect to the customs classification 6-digit Harmonized System (HS6). Each HS6 chapter is the sum of one or more sub-NC8s.

[1] The HS classification at 6-digits level, developed by the World Customs Organization (WCO), is revised every 5 years (last time in 2017).

United Nations - db Monthly Comtrade

Since 2010 the United Nations Statistics Division has been carrying out an experimental project called Monthly Comtrade, for the construction of an online database, accessible at http://comtrade.un.org/monthly/, on monthly foreign trade flows reported by UN countries.

Data are available in dollars and kilograms, for each product of the Harmonized System customs classification at 6-digits level (HS6).

US Census Bureau - db UsaTrade

US Census Bureau disseminates monthly foreign trade data reported by the United States of America through the portal UsaTrade Online (https://usatrade.census.gov/), according to the Harmonized System customs classification at 10-digit level (HS10). HS10 consists of a further specification with respect to the customs classification 6-digit Harmonized System (HS6). Each HS6 chapter is the sum of one or more sub-HS10s.

Estimation of Year-End

Foreign trade data are processed by StudiaBo with the objective to support the analysis also of more recent events. As already mentioned, however, the Statistics Division of the United Nations poses no terms for the transfer of information and therefore trade flows regarding previous year can often be lacking. With the aim to overcome this drawback, StudiaBo has implemented an estimation procedure to *close the year* on the basis of the monthly declarations available. This estimation procedure consists, first, isolating the last two years elapsed. So, from the *n* months already subject to reporting, for each flow (country of origin - country of destination) change rates were calculated for values and quantities. This operation allows to estimate the information of the foreign trade of *12-n* months still missing and hence to derive a first hypothesis of year-end: the StudiaBo procedure, in fact, applies to preceding annual UN Comtrade figures the dynamics expressed by the different monthly databases. Plase note that, in the case of non-EU countries that do not adhere to Monthly Comtrade, the estimation procedure appears quite *imposing*. Given a flow between a country of origin and a country of destination, the procedure detects the following cases:

- in the case of the construction of a closed sample for the last two years, if it is possible to calculate a rate of change for both exporter and importer, the flow not yet explicitly stated is calculated applying a change rate equal to the average between the 2;
- in the case the information available enables to derive only a change rate either of the exporter or of the importer, the flow not yet explicitly stated is calculated through the characteristic ratio (rate of change) available for the most recent year.
- in the case it is not possible to calculate any rate of change for a given flow, the latter is calculated applying the ratio of the total exports of the product last year history and the total exports of the product in the penultimate year.

Normalization

As mentioned before, information on foreign trade are characterized by a *double statement*: each flow is, in fact, declared by both the exporting country and the importing country, with possible margins of difference. As known, the values declared by the importer are affected by the incidence of the costs of insurance and freight (CIF: Cost, Insurance and Freight) and are therefore generally higher than the flows declared by the exporter (FOB: Free On Board). In this regard, the procedure StudiaBo, based on the *Mirror Statistics* methodology, detects and manages the following types of cases:

- the trade flow is declared only once, by the exporter or the importer (eg, for reasons of commercial secrecy or small economic importance of the flow). In this case, StudiaBo estimates the missing declaration, considering a difference in value between statements at CIF prices and FOB prices equal to 5 %, in agreement with a broad empirical literature that estimates economic costs of transport and insurance as a factor proportional to the total value of between 3 and 8%;

- the trade flow is declared twice, both by importer and exporter, and the two statements are mutually consistent [2]. In this case, no adjustments are made by the procedure.

- the trade flow is declared both by importer and exporter, but the two statements are mutually inconsistent. In this case, the StudiaBo procedure recalculates the value of the flow at FOB and CIF prices, using the following simple average:

$$FOB = (V_{exp} + V_{imp}/1.05)/2 \qquad (1)$$

$$CIF = (V_{exp} * 1.05 + V_{imp})/2 \qquad (2)$$

where V_{exp} e V_{imp} represent, respectively, the value declared by the country of origin and the value declared by the country of destination.

The technique just described allows to derive higher level information from the data available, particularly with regard to the value measurement. The measurement of quantities, in contrast, requires an additional step, through the identification of outliers and the estimation of the missing data, which will be described in the next chapter.

[2] StudiaBo consider a double statement as consistent whenever the two declarations do not differ in value by more than 20%.

Outliers

One possible cause of discrepancy between statements of import and exports in quantity is the presence of outliers or measurement errors which, if not treated, can affect the understanding of the economic phenomenon. Quantity declarations, whether expressed in kilograms or in supplementary unit measure, are therefore subject to some control filters. First, StudiaBo has decided to eliminate the information in quantity estimated by the Statistics Division of the United Nations, considering it more appropriate to refer to a well codified internal methodology. Therefore, on the basis of the available declarations, the following key ratios are constructed:

- an average unit value per flow, given by the ratio of monetary value and kilograms;
- a conversion factor, given by the ratio of kilograms and supplementary unit measure.

On the basis of the orderly distribution of these key ratios, a *range of reliability* is defined (based on the first and ninth decile of the distribution) within which the observations are considered to be reliable. Observations excluded from this range are therefore recalculated to be in the range of validity, resulting in a *smoothing* of the distribution. This method is used for all traded goods that have been identified previously as *differentiable*[3]. In case, however, the preliminary analysis of differentiability has concluded the substantial homogeneity of a good, the range of reliability of the key ratios is restricted around the median.

[3]To the purpose of distinguishing *commodity goods* from *qualitatively differentiable goods*, StudiaBo has built an econometric panel data model at fixed effects, through which it is estimated the elasticity of export shares to price changes.

Missing Data

In cases where the quantity information are not available or estimated by the United Nations Statistics Division, StudiaBo is committed to rebuild the data on the basis of the following internal method of interpolation:

- First it is built the time series of annual average unit values per single product, given by the characteristic ratio between values and kilograms of the product for each year). This series represents the base dynamics followed by the prices of the product object of analysis and is characterized by one-dimension (time). Note that in the case little information is available for the last year elapsed, the characteristic ratio is set equal to that of the penultimate year. If, on the contrary, the information available are sufficiently consistent, the average unit value is calculated on the basis of the closed sample of observations in the last two-year period.

- If it is not possible to give an annual characteristic ratio to a product, StudiaBo makes a first interpolation using as a reference the years available.

- Given the time series of annual average unit values, a 2-dimensional matrix is built, where the price drivers is not only the year but also the exporting country. Note that, for reasons related to the robustness of information, the values within that matrix are made explicit only for those countries that record in a base year (e.g. 2010) a share on the world trade of the product of at least 1%. In the case of less relevant exporting countries, the characteristic ratio will coincide with that of the annual time series of the product. In the case that, in a given year and for a given exporter, information in quantity is not available, StudiaBo's procedure will again recourse to interpolation, using the index given by the characteristic ratio for exporter and year compared to the characteristic ratio for year.

- Finally, the Cartesian ratio of observations is built, which identifies, for the total time period, all the possible combinations of flows between a country of origin and a country of destination. For every combination of exporter and importer, an interpolation is carried out of the index given by the characteristic ratio for flow and the characteristic ratio for exporter.

- Once reached the highest level of completeness of the available data, the information in kilograms is reconstructed from the average unit values, obviously only for flows that have a counterpart in value.

- After the reconstruction of the information in kilograms, this process is repeated for a second type of characteristic ratios, i.e. the conversion factor, which explicit for each product the weight per supplementary unit measure.

Price Ranges

In order to analyse the quality characteristics of the international trade flows of a product, the StudiaBo's procedure gives each flow a variable (*Price Range* (R), built on the basis of the ordered distribution of average unit values (the ratios between values and kilograms). In detail, the values of international trade are deflated by a production price index, which is built on-purpose for about 200 supply chains. This price index summarizes the impact of the following inputs:

- the cost of labour;
- the price of materials.

In particular, the cost of labour is calculated from the hourly labour costs per country, provided by the U.S. Bureau of Labor Statistics. The average hourly labour cost is weighed on the basis of the export levels of the more than 150 countries of Ulisse classification and then adjusted by a fixed percentage attributable to productivity changes. In turn, the price of materials incorporates both the oil price (as a proxy of energy consumption) and the price of the main commodity intervened in the production process of the supply chain. By deflating average unit values it is possible to build a distribution of prices comparable to each other, as relativized to a base year. Once identified the deciles of the distribution, these are used as thresholds for identifying different price ranges. Then it is possible to assign a dynamic to these price ranges: for each year, in fact, the specific deciles of the distribution are calculated, by applying to the deciles of the base year (e.g. 2010) the change rate embedded in the production price index, according to the equation:

$$Decile_{n,year} = Decile_{n,2010} * Indice_{year}/100 \quad (3)$$

where $Decile_{n,year}$ represents the decile that in the given year holds the position *n*, $Decile_{n,2010}$ is its counterpart calculated in the base year (2010) and $Indice_{year}$ represents the value expressed by the production price index in the given year. Therefore, the price ranges are dynamically identified to distinguish the following ranges of price (quality):

- *LL -Low Low*: identifies the low-quality market segments, comprehensive of all the flows with an average unit value below the second decile of the orderly distribution of prices at which the product is marketed worldwide;
- *LM - Medium Low* identifies the medium-low quality market segments, comprehensive of all the flows between the second and fourth decile of the orderly distribution of prices at which the product is sold worldwide, between the fourth and sixth decile, in which the product is marketed; item *HM - High Medium* identifies the medium-high quality market segments, comprehensive of all the flows between the sixth and eighth decile of the orderly distribution of prices at which the product is sold worldwide;
- *HH - High High* identifies the high-end quality market segments, comprehensive of all the flows with an average unit value higher than the eighth decile of the orderly distribution of prices at which the product is marketed worldwide.

This process of attribution, of course, applies only in cases where the product has been previously identified as *differentiable*, through the estimation of the apposite economic model. In the case of *not*

differentiable goods, such that an exporting country is impossible to extract a significant premium-price from the market, in the construction of the variable Price Range (R), the standard price range MM (Medium-Medium) is applied to all flows of that product.

Quantity at constant prices

As mentioned before, the main measures of foreign trade flows concern values and physical units (generally expressed in kilograms), whose ratios are the so-called *average unit values*, often regarded as a proxy of prices. Average unit values are widely used in the procedure for the construction of DW Ulisse. It should however be pointed out that changes in average unit values, besides reflecting price changes, might also incorporate changes in product quality. In order to develop analyses that consider those amendments, StudiaBo preferred to unbundle the qualitative effects on the prices, to include them in a new variable (Q), which measures quantity at constant prices. By doing so, it is possible to obtain a price indicator, based on the ratio between values and quantities at constant prices, more *informative* because not vitiated by qualitative modifications. In detail , the variable Q is constructed for a single foreign trade flow with reference to a base year (e.g. 2010), isolating the quantity on the basis of the price range at which the flow belongs for that year. For that year the procedure builds both the average unit value for each specific combination [country of origin - country of destination - price range] and the average price for each price range. In the case, for a given flow, there is a qualitative increase compared to 2010, such as to be associated to a different price range, then the average unit value of 2010 is abandoned in favour of the specific average price associated with the new price range of reference. This causes a consequent increase of the quantity expressed at constant prices (Q), as explained by the following equation:

$$Q_{flow,year} = K_{flow,year} * AUV_{2010} \qquad (4)$$

where $Q_{flow,year}$ identifies the quantity at constant prices in any given year for a given combination [country of origin - country of destination - price range], $K_{flusso,year}$ identifies the kilograms object of foreign trade flows in the given year for a given combination [country of origin - country of destination - price range], AUV_{2010} represents the average unit value in 2010 related to the price range associated to the flow.