Methodology for the detection of outliers in quantities on Trade Map yearly time series and for subsequent estimation of quantities

31 August 2016

To improve the quality of Trade Map and facilitate data analysis, a method has been elaborated to detect outliers in quantities among trade in goods data. Many reasons can explain the presence of outliers, one of which being a recording error. The objective of the outliers detection in Trade Map is to inform users that some data have to be analysed carefully.

Tested data

To ensure that tests are not affected by a change in methodology, the following filters are applied.

- Only direct data are tested.
- Only product codes at HS 6-digit level are tested
- Only time series of products which are not affected by revision changes of the Harmonised System (neither in 2002, 2007 or 2012) are tested. In other words, only the products which have not been revised since 1996 are tested.
- Only series which unit of quantity never change over time are tested.
- Only series having less than 20% of missing data (quantity=0) are tested.

Quantities are tested within yearly time series for each triplet {reporting country, partner country, product}.

In Trade Map, a quantity is tagged as an outlier only if the quantity and the related unit value are both detected as outliers.

Conducted tests

Two modified Z-tests are applied:

- one for each quantity:

$$\left| 0.6745 \times \frac{(q_{ijky} - Me(Q))}{Me(|q_{ijky} - Me(Q)|)} \right| > 3.5$$

where:
_ i, j, k and y are respectively the reporting country, partner country, product and year
_ q_{ijky} is the tested quantity
_ Q is the series of quantities for each triplet {reporting country, partner country, product}
_ Me is the median function
_ |…| is the absolute value operator
_ 0.6745 and 3.5 are commonly used constants for modified Z-tests (see for instance Alcaraz Garcia, 2010 and Iglewicz & Hoaglin, 1993).

- one for each unit value:

$$\left| 0.6745 \times \frac{(Ln(uv_{ijky}) - Me(Ln(UV)))}{Me(|Ln(uv_{ijky}) - Me(Ln(UV))|)} \right| > 3.5$$

where:
_ uv_{ijky} is the tested unit value (in US$ per unit of quantity)
_ UV is the series of unit values for each triplet {reporting country, partner country, product}
_ Ln is the natural logarithm function

Natural logarithm is applied to unit values so that particularly low unit values can be detected as outliers. Indeed, in certain cases, a unit value can stay positive while being very close to zero (e.g. 10^{-9} US$ per ton). Absolute deviation to the median unit value can be rather small but impact further computations substantially. Performing logarithm transformation on unit values before applying modified Z-test takes these cases into account.

If both tests are true, i.e. if both quantity and unit value are detected as outliers, then they are both estimated as detailed below.

Estimation of unit values and quantities

If estimations are required, unit values are firstly estimated. For one year, one reporting country and one product, if the sum of values in US$ from partners without quantity outliers represents more than 50% of the value for the partner “world”, the average unit value of partners is used for the selected year to calculate the quantity.

For one year, one reporting country and one product code, if the sum of values in US$ of partner countries without quantity outliers represents less than 50% of the value for the partner “world”, then the unit value is estimated by the average of unit values between the previous year and the following year for the same partner country. If the quantity outlier is the last year available, then the average unit value of the two previous years is calculated. If the quantity outlier is the first year available, then the average unit value of the two following years is calculated.

Once unit values have been estimated, quantities are recomputed as follows:

$$q^*_{ijky} = \frac{v_{ijky}}{uv^*_{ijky}}$$

where:

_ q^*_{ijky} is the newly estimated quantity

_ uv^*_{ijky} is the newly estimated unit value

_ v_{ijky} is the corresponding trade value (in US$)

The quantity for the partner “world” is always the sum of quantities from individual partners. For this reason, if a quantity outlier is identified for the partner “world” and not for individual partner countries, the unit value is estimated for the partner “world” and applied to all partner countries. Quantities are then recomputed according to the same formula as before using this newly applied unit value.

Estimated quantities and unit values are colored in dark green in Trade Map.

Iterative process

Since time series can potentially have more than one outlier, this procedure is run a second time so that outliers that would not be detected during the first run are replaced by an estimation following the same procedure as before. One notable exception: during this second run, if a quantity outlier is identified for the partner “world” the unit value is estimated for the partner “world” and systematically applied to all partner countries.