Agricultural trade data processing

FAO, in its FAOSTAT corporate statistical database, provides access to the detailed agricultural trade data of many countries of the world. These trade datasets have been standardized by the FAO Statistics Division and are available by quantities and values, imports and exports, reporter and partner country. They are available at different disaggregation levels within the commodity classification system of Harmonized System (HS)[[1]](#footnote-1). National FBS compilers are encouraged to make use of these trade datasets for the production of their FBS. More generally, countries are encouraged to compare their national agricultural trade with information from other countries, to study regional and global patterns, and importantly, to make use of trading partner “mirrored” data to cross-check national figures, or to fill-in missing data. Furthermore, these datasets can be used to adjust national data, where required, based on the commodity unit-value concept.

### UNSD Tariff-line dataset

The Statistics Division of FAO receives trade data for most countries from the [United Nations Statistical Division](http://unstats.un.org/unsd/default.htm) which collects the data from the countries and stores them in the [Commodity Trade Statistics Database UN Comtrade](http://unstats.un.org/unsd/comtrade_announcement.htm). These datasets are as originally reported by the national customs offices (usually) and are by detailed flows (as described in the above paragraph). The level of commodity detail is country-specific, with some countries reporting at the basic 6-digit level of the HS, while others go up to the 12-digit HS detail.

The Harmonized Commodity Description and Coding System (HS) is maintained by the World Customs Organization. The global coverage, by value, of this dataset is more than 90 percent, in any given year. A typical non-standardized country trade dataset, at 10-digit HS level, would look like outlined in ():

Table 1: Basic trade information by commodity, partner and reporter

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Reporter | Partner | HS code | Flow | Weight | Quantity | Q-unit | Value |
| 2011 | 842 | 484 | 8501314000 | 1 |  | 33498070 | 5 | 366057280 |
| 2011 | 842 | 484 | 7321190040 | 1 |  | 242 | 5 | 21321 |
| 2011 | 842 | 591 | 7326908588 | 2 | 1896 | 1896 | 8 | 13011 |
| 2011 | 842 | 792 | 6104499060 | 1 | 829 | 232 | 11 | 78839 |
| 2011 | 842 | 699 | 2907191000 | 1 | 20000 | 20000 | 8 | 128598 |

Reporters and trade partners are represented with three-digit numerical [codes](http://unstats.un.org/unsd/methods/m49/m49.htm) used by the Statistics Division of the United Nations, based on the international standard M49 codes. The “flows” indicate import (“1”), export (“2”), re-import (“3”) and re-export (“4”). The “weight” is in kilograms, while the “quantity” indicates other units of measurement such as “Number of items” or “Volume in liters”. The full list of units and their descriptions is in Annex I of Quantity and Weight Data in UN Comtrade[[2]](#footnote-2). The commodity codes, as mentioned above, follow an internationally comparable standard at the 6-digit level, but become country-specific at the higher digit levels (7 digits and above)[[3]](#footnote-3). Below () is an extract of the HS classification at 10-digit level reported by a country:

Table 2: Example of trade flow disaggregation at the 10 digit level

|  |  |
| --- | --- |
| HS | Description |
| 02 | CHAPTER 2 MEAT AND EDIBLE MEAT OFFAL |
| 0202 | Meat of bovine animals, frozen: |
| 0202.10 | * Carcases and half-carcases: |
| 0202.10.05 | * + Described in general note 15 of the tariff schedule ...: |
| 0202.10.05.10 | * + - Veal |
| 0202.10.05.90 | * + - Other |

## Key processing steps

The process outlined below describes the key processing steps employed by FAO to prepare trade data sets for dissemination and importantly for use in the SUA/FBS process. FBS compilers at country level may want to critically review these steps, evaluate their suitability for their domestic data and purposes and either develop their own trade data processing system or simply take advantage of the data made available by FAO.

The FAO processing of trade data is organized around the following steps. To obtain a quick overview, FSB compilers may also want to refer to the flow chart depicted in below.

### Pre-validation

In a first step, a series of programmes to identify missing quantity data is run and based on these, unit-value based outlier validations are performed. These checks ensure that the data meet the minimum requirements for further processing. They also provide a first opportunity to mend such data problems and avoid that low quality information is maintained throughout the rest of the data processing process.

### Dataset Standardization

Dataset standardization includes a number of additional preparatory steps to facilitate further data processing. These include basic formatting steps (e.g. unifying the two quantity columns into one column), unit of measurement conversion (e.g. all quantities are converted to metric tonnes, numbers, etc.), links of country codes to country names; and the conversion from HS to FCL/CPC[[4]](#footnote-4). If/when successfully completed, these steps will produce the FAO standardized trade dataset.

The above steps are applied to all reporting countries. Once completed, we can also calculate regional and global unit-values (medians) for every commodity and for a specific year.



Figure 1: Trade processing steps at FAO

### Validation of trade data

Based on the results of the missing data and outlier report, further initial validation steps (pre-validation) can be undertaken. These include:

1. Estimate missing quantities (and very rarely the monetary values) based on the reporter trade flow median unit-value.
2. Adjust quantity outliers (out-of-range values) again based on the reporter trade flow median unit-value (often, quantity errors can be due to incorrect decimal points/commas).
3. When a significant unit-value median cannot be calculated in the above two cases, due to an insufficient number of trade flow observations in the reporter dataset, use is made of the continental/regional or global median unit-values for that commodity instead.
4. In some cases, the unit-value of the “mirrored” trading partner flow may be used. However, since imports are reported on a “free-on-board”(FOB) basis, while exports are on a “cost-insurance-freight“ (CIF) basis, an FOB/CIF adjustment factor must be used. This factor, which reflects the difference in FOB and CIF values, is commodity-specific and dependent on the mode of transportation and geographical distances. “Mirrored” partner data are corresponding flows of imports or exports as declared by the partner country; needless to say that they are specific to a specific commodity and year.

The above 4 steps warrant a more detailed description so that they can be repeated or modified by national FBS compilers. Obviously, also the more detailed description will off with the outlier detection process.

#### Outlier detection

Unit value (UV=monetary value/quantity) outliers, as observations located outside the range, are identified using this formula:

where and are the lower and upper quartiles respectively (25% to 75%, with the median at 50%), and is a non negative constant. Here we use = 1.5, as was suggested by John Tukey[[5]](#footnote-5) in his basic approach for detection of potential outliers. Such an approach is more outlier-robust, than the arithmetic mean. The chart below () shows the combination of missing quantities and unit-value outliers for reporting countries in a given year:

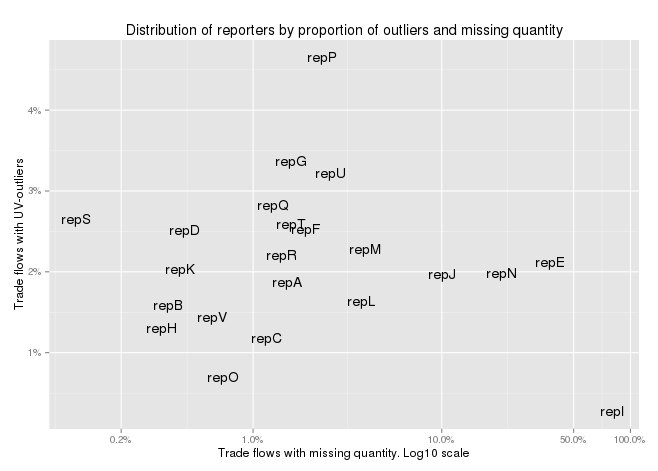


Figure 2: results of outlier tests

#### Imputation of missing quantities/values and replacement of outliers

Once the outlier detection process is completed, the next steps are to find criteria to remove or keep outliers and impute missing quantities.

As an example (), our validation checks show that the data reported by country X, in a particular year, in HS chapters 2, 10 and 15, have 17 trade flows with missing quantity and 33 trade flows with UV-outliers. The reported partners are Y1 to Y4.

Table 3: identifying and mending outliers, based on trade flows with missing quantities and outlier unit values

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Reporter | Partner | Flow | Commodity  (FCL code) | Quantity | Value | UV | UV-me |
| X | Y1 | Import | 271 | - | 9193 | - | 3.71 |
| X | Y1 | Import | 271 | 89 | 3267 | 36.71 | 4.91 |
| X | Y2 | Import | 1035 | - | 194271 | - | 4.76 |
| X | Y3 | Import | 56 | 23 | 2937 | 127.7 | 3.84 |
| X | Y4 | Import | 870 | - | 112727 | - | 4.49 |
| X | Y1 | Import | 79 | - | 2479 | - | 1.88 |
| X | Y1 | Import | 258 | 192 | 4382 | 22.82 | 1.69 |
| X | Y1 | Import | 1243 | 2853 | 24030 | 8.42 | 3.87 |
| X | Y1 | Import | 89 | - | 7833 | - | 1.24 |
| X | Y1 | Import | 340 | 769 | 22842 | 29.7 | 3.66 |

(for simplicity, not all the trade flows used to calculate the median unit values are shown)

We can now impute a missing quantity, or adjust an outlier, using the reporter median unit value for the respective commodity.

Table : trade flows with adjusted quantities

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Reporter | Partner | Flow | Commodity  (FCL code) | Quantity | Adjusted  quantity | Difference |
| X | Y1 | Import | 271 | - | 2478 | - |
| X | Y1 | Import | 271 | 89 | 665 | -576 |
| X | Y2 | Import | 1035 | - | 40813 | - |
| X | Y3 | Import | 56 | 23 | 765 | -742 |
| X | Y4 | Import | 870 | - | 25106 | - |
| X | Y1 | Import | 79 | - | 1319 | - |
| X | Y1 | Import | 258 | 192 | 2593 | -2401 |
| X | Y1 | Import | 1243 | 2853 | 6209 | -3356 |
| X | Y1 | Import | 89 | - | 6317 | - |
| X | Y1 | Import | 340 | 769 | 6241 | -5472 |

## 

#### Imputation using “mirrored” data from a trading partner

In an additional step, we can now make use of partner data and impute missing information through “mirroring”. Before so doing, we first need to validate or reject the possible outliers. In our example, the validation algorithm identifies the following trade flows as having outlier unit values:

Table : Updated trade matrix with validated data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Reporter | Trade partner | Flow | Commodity  (FCL code) | Quantity | Value | Unit Value | UV median |
| 2011 | X | **Y5** | **Import** | 1223 | **160571** | 4932146 | 30.71 | 5.83 |
| 2011 | X | Y6 | Import | 994 | 105784 | 2191650 | 20.72 | 7.15 |
| 2011 | X | Y7 | Export | 1064 | 726 | 11780 | 16.23 | 4.27 |

In this case, the *import* quantity outlier will be adjusted using the *export* “mirrored” trading partner data. Let's describe in more detail, how this “mirroring” is done with first row from table above.

Now we check the mirrored data as reported by the partner Y5:

Table : Step-by-step introduction into the “mirroring” process

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Reporter | Trade partner | Flow | Commodity  (FCL code) | Quantity | Value | Unit Value | UV median |
| 2011 | Y5 | **X** | **Export** | 1223 | 131206 | 2665550 | **20.32** | 5.07 |

The corresponding export unit value is calculated based on the corresponding quantity and value data:

Partner export Value/Partner export Quantity = Partner export UVF

266550/131206 = 20.32

However, since usually exports are reported on a “free-on-board”(FOB) basis (which excludes additional costs such as transport, insurance, etc.), while imports are on a “cost-insurance-freight“ (CIF) basis, the export UV is thus expected to be lower than the import UV. To adjust for this, an FOB/CIF adjustment factor is being used. This factor, as mentioned earlier, is commodity-specific and dependent on the mode of transportation and the relevant geographical distances, among others. While specificity is always important, there is also a number of studies that provide standard factors for adjustment, broadly applicable to most commodities and trading partners. Let us assume, for this example, that the FOB/CIF factor here is 1:1.15 (i.e. 15%).

Therefore, the final UV from the above calculation is:

20.32 x 1.15 = 23.34

Now we apply this adjusted partner export UV to estimate the reporter import quantity:

Reporter import total value/Partner adjusted export UV = Adjusted reporter quantity

4932146 / 23.34 = 279132.3

Table 7: Introducing CIF/FOB corrections to the mirrored trade dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Reporter | Trade partner | Flow | Commodity  (FCL code) | Mirrored UV | Value | Corrected quantity |
| 2011 | X | Y5 | Import | 1223 | 23.34 | 4932146 | **279132.3** |

#### Using “mirrored” trade data for non-reporting countries

Not all the countries of the world provide their trade data to the United Nations Statistical Department. (UNSD). Thus, also the trade dataset obtained by FAO from the UNSD does not have a full global coverage. For these non-reporting countries we have to resort to other sources and methodologies – mainly mirrored trade data - to obtain or estimate their respective trade figures. To identify those countries that are missing as reporters in our fully validated and processed dataset, we run a check using the full FAO country codes list.

To this end, we use the global trade flow tables to fill-in the missing imports and exports of non-reporting countries. Briefly, the steps followed are:

In a first step, we identify/filter for each non-reporting country all the reported trade flows which indicate this non-reporter as a “partner” in our validated dataset. In the example table below, we have a simplified “filter” for non-reporting partner Y5, showing only two commodities which are reported as being traded with Y5 globally in that specific year ().

Table 8: starting table for mirroring flows after validation and imputation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | All  Reporters | “non-reporting”  Trade partner | Flow | Commodity  (FCL code) | Quantity | Value |
| 2011 | X | **Y5** | Import | 1223 | 160571 | 4932146 |
| 2011 | A | **Y5** | Import | 1223 | 80120 | 2301650 |
| 2011 | C | **Y5** | Export | 1064 | 726 | 11780 |
| 2011 | D | **Y5** | Export | 1064 | 2000 | 35780 |

In a second step, we undertake 3 distinct operations:

(a) The *exports* become the *imports* of the non-reporting country for each specific commodity; and vice-versa.

1. The quantities are taken as they are; whilst the monetary values are adjusted for FOB/CIF, as previously described.
2. The non-reporting country is now listed as a “reporter”, and the reporters become partners”.

Table 9: Final trade matrix after mirroring

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Reporter | Partner | Flow | Commodity  (FCL code) | Quantity | Value |
| 2011 | **Y5** | **X** | Export | 1223 | 160571 | 4120350 |
| 2011 | **Y5** | **A** | Export | 1223 | 80120 | 2110443 |
| 2011 | **Y5** | **C** | Import | 1064 | 726 | 13810 |
| 2011 | **Y5** | **D** | Import | 1064 | 2000 | 41455 |

With this we now have the **complete global trade flow dataset** (also see ) which will be disseminated through the FAOSTAT website.

## Trade flow aggregation and commodity balances

All the trade flows are now aggregated by reporter, commodity and flow (import/export) for that year. So, the trade flows in the table above would be aggregated as below ():

Table 10: Aggregated, total flows for dissemination

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | Country | Flow | Commodity  (FCL code) | Quantity | Value |
| 2011 | Y5 | **Export** | 1223 | 240691 | 6230793 |
| 2011 | Y5 | **Import** | 1064 | 2726 | 55265 |

This “totals” dataset by commodity, by country, by region and world, exists alongside the original validated trade flow dataset, and allows for the manual input of commodity/country data from other sources; and shows the global totals by commodity per year (to check the import vs. export commodity balances).

The other sources used to obtain trade data for non-reporters, or to adjust reported data, mainly include official statistical publications (usually online), data obtained from specialized commodity institutes (such as OilWorld, International Coffee Organization, etc.), and food aid statistics. Presently, data from such sources are entered manually into the system; however, in the near future, these data will also be “harvested” in an automated manner (similar to the download of data from UNSD).

This **final data set with the global “totals”** (including data from other sources, and commodity imbalances minimized) is also disseminated through the FAOSTAT website. National FBS compilers can use extractions from these datasets for the compilation of their national Food Balance Sheets. How to integrate trade into the FBS dataset has already been explained in Chapter 3.

1. While only flows at HS6 level follow the same classification across all countries, the FAO Statistics Division plans to make available also flows at higher level of disaggregation (tariff line level). [↑](#footnote-ref-1)
2. <http://unstats.un.org/unsd/tradekb/Knowledgebase/Quantity-and-Weight-Data-in-UN-Comtrade> Quantity and Weight Data in UN Comtrade [↑](#footnote-ref-2)
3. [http://madb.europa.eu](http://madb.europa.eu/) Descriptions of country-specific HS-codes are provided by Market Access Database and copyrighted by Mendel Verlag, Germany. [↑](#footnote-ref-3)
4. FCL is the “FAO Commodity List” and CPC is the “Central Product Classification”. FAO uses the CPC version 2.1 “expanded”: <http://www.fao.org/economic/ess/ess-standards/commodity/en/> [↑](#footnote-ref-4)
5. J.W. Tukey. Exploratory Data Analysis. Addison-Wesley, 1977. [↑](#footnote-ref-5)