faoswsTrade

Marco Garieri
Food and Agriculture Organization
of the United Nations

Abstract

This vignette provides a detailed description of the various data sources and procedures used in the trade modules.

Keywords: Agricultural Trade, Tariff Line, Eurostat, Mirroring.

The trade module is divided in two submodules: `complete_tf_cpc` and `total_trade_CPC`. Each module is year specific. This means that, at the time being, the trade module run indipendently for each year. In order to run the `total_trade_CPC`, the output of `complete_tf_cpc` is needed.

1. Complete tf cpc

1.1. Data

Raw data are provided by the SWS Team (subunit of Team F) for both UNSD Tariffline and Eurostat Data. The data have been already prefiltered:

Eurostat
- code of reporter (declarant) just numeric (letters are not allowed)
- code of partner (partner) just numeric (letters are not allowed)
- code of CN8 (product_nc) just numeric (letters are not allowed)

UNSD
- code of HS (comm) just numeric (letters are not allowed)

The module downloads only records of commodities of interest for Tariffline Data. The HS chapters are the following: 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 33, 35, 38, 40, 41, 42, 43, 50, 51, 52, 53. In the future, if other commotidy are of interest for the division, it is important to include additional chapter in the first step of the downloading. For Eurostat Data no filtering is applied.

1.2. Process

Mapping UNSD Tariffline and Eurostat data

At this stage a standardization/mapping step is performed. The details are devided between UNSD Tariffline and Eurostat due to the nature of the differences among the two datasets.

UNSD Tariffline UNSD Tariffline data reports area code with Tariffline M49 standard (which are different for official M49). The area code is converted in FAO country code using a specific convertion table provided by Team ENV. Area codes not mapping to any FAO country code or mapping to code 252 (which correpond not defined area) are separately saved and removed from further analyses.
Commodity codes are reported in HS codes (*Harmonized Commodity Description and Coding System*). The codes are converted in FCL (FAO Commodity List) codes. This step was performed using the package (`hsfclmap` developed by Alexander Matrunich), but it is now incorporated in the module. In this step, all the mapping between HS and FCL code is stored. The algotithm tries to map at all possible lenght (i.e. if a reporting country has a record with hs code at 12 digits and in the package, for the same reporting country, a HS-to-FCL mapping is available at a lower level, for example 10 digits, the algorithm will include in the mapping all the records having the same 10 digits).
If a country is not included in the package of the mapping for that specific year, all the records for the reporting country are removed.
All records filtered out are saved in specific variables, but not used in the output at the time being. Further develop of the module Information of the FCL units is added at the end of this step.
Just for UNSD Tariffline data convertion of units of measurements are applied to meet FAO standards, where all weights are reported in metric tonnes, animals in heads or

1000 heads and for some commodity just the value is provided.
The flow codes of re-Import (4) are recoded into Import (1) and codes of re-Export (3) to Export (2). This procedure is applied following UNSD standards:

> Distinction between Exports and Re-exports / Imports and Re-imports
> Exports of a country can be distinguished as exports of domestic goods and exports of foreign goods. The second class is generally referred to as re-exports. The exports shown in our database contain both the exports of domestic and foreign goods. Re-exports are exports of foreign goods in the same state as previously imported; they are to be included in the country exports. It is recommended that they be recorded separately for analytical purposes. This may require the use of supplementary sources of information in order to determine the origin of re-exports, i.e., to determine that the goods in question are indeed re-exports rather than the export of goods that have acquired domestic origin through processing. Re-imports are goods imported in the same state as previously exported. They are included in the country imports. It is recommended that they be recorded separately for analytical purposes. This may require the use of supplementary sources of information in order to determine the origin of re-imports, i.e., to determine that the goods in question are indeed re-imports rather than the import of goods that have acquired foreign origin through processing. There are several reasons why an exported good might return to the country of origin. The exported good might be defective, the importer might have defaulted on payments or cancelled the order, the authorities might have imposed an import barrier, or demand or prices in the country of origin might have made it worthwhile to bring the good back.

Eurostat Eurostat data reports area code with geonomenclature standard. The are code is converted in FAO country code using a specific convertion table provided by Team B/C. Area codes not mapping to any FAO country code or mapping to code 252 (which correpond not defined area) is reported and the records for these area codes are removed.
Commodity codes are reported in CN8 codes (Combined Nomenclature 8 digits). The codes are converted in FCL (FAO Commodity List) codes. This step is performed using the same package (`hsfclmap`) as for UNSD Tariffline. If a specific record has a CN8 code not mapping to any specific FCL code, then the record is reported and removed.
If a country is not included in the package of the mapping for that specific year, all the records for the reporting country are removed.

The possible solution for the missing links in the future using Natural language processing routines to read the metadata.
Information of the FCL units is added at the end of this step.
Values are converted from EUR to USD using the table with avarage currency for each year provided by Team B/C.
Eurostat data are already provided in the correct units of measurements and do not need futher conversions.

Unified Official Trade Flows Dataset

UNSD Tariffline and Eurostat datasets are ready to be merged.

Standardization, editing and outlier detection

- **Application of Notes** Perennial and yearly specific notes are mdb files provided by the Team B/C already saved in a R friendly dataset. The notes might be year specific or for

all years (in this case reported as NA) and might refer to HS or/and FCL codes. This notes (or adjustments) were developed during the years and they are available from 1997 to 2013. Notes of 2014 are copied from notes in 2013, as a partial solution, but this need future work in the future.
Comparing results between the new and the old procedure showed that sometimes the discrepancies between the two results are due to the application of the notes.
Remark: at the time being, the notes with unspecified year and with application of a factor 1000 are removed, since in the previous years UNSD was reporting some data in tonnes, while now it reports all data in kg.

- **Unit Values computation** For each record having both quantity and value (thus excluding all commodity reported just as value), the unit of value (u_v) is computed as following:

$$u_v = \frac{qty}{value} \tag{1}$$

- **Outlier Detection and Imputation** The outlier are calculated based on the distribution of the unit of value for the same country, year and flow at the HS level. The reason to identify the outlier at the HS level is due to the fact that, under the same FCL code, different commodity might fall (i.e. maize seed and seed). The outlier are detecting using the Tukey's procedure:
  - The Tukey's five number summary are calculated: minimum (m), lower-hinge (lh), median (med), upper-hinge (uh) and maximum (M).
  - The coefficient for the outlier detection is set up as suggested by Tukey to 1.5 ($coef$).
  - For each value is calculated a specific distance from the lower or the upper-hinge in the following way:

$$\text{x is outlier if } \begin{cases} x < lh - coef * iqr, & \text{lower outlier,} \\ x > up + coef * iqr, & \text{upper outlier.} \end{cases} \tag{2}$$

  where iqr is the interquartile range.

  The outlier are then corrected using the corresponding value and dividing it by the median unit of value of that specific commodity, country, flow and year. In this way only few official data are corrected.
  Remark: in the module, one of the input parameter for the user is the oultlier coefficient By default this is set up to 1.5. More info regaring the outlier coefficient is given in the Future Work section.

- **Missing Quantities Detection and Imputation** For records in which the commodity has to be reported in quantity and the quantity is missing and the value is present, the corresponding quantity is imputed dividing the corresponding value by the median of the units of value of the corresponding commodity (HS level/country/flow/year)

Mirroring and Balancing

The module produce the list of non-reporting countries: these are the countries present as partners but absent as reporters. For this countries the mirroring routine is applied: the corrisponding trade of the non-reporting countries are extracted from the partners inverting the flows. The quantities are the same while the values are corrected by a factor of 12% due to the

CIF/FOB conversion. This need more work, details in the Future Work section.

1.3. Flags

Both records with imputation with outlier or mirroring imputation have a special flag:

- `flagObservationStatus`: this flag is **I**, which means imputed
- `flagMethod`: this flag is **e**, which means estimated.

For all the other records empty string flags are saved.
More information on the Flag is given in the Future Work section.

1.4. Convertion to FAO SWS standards

At this point the table is almost ready to be save in the SWS. Additional mapping are necessary in order to respect the SWS standards:

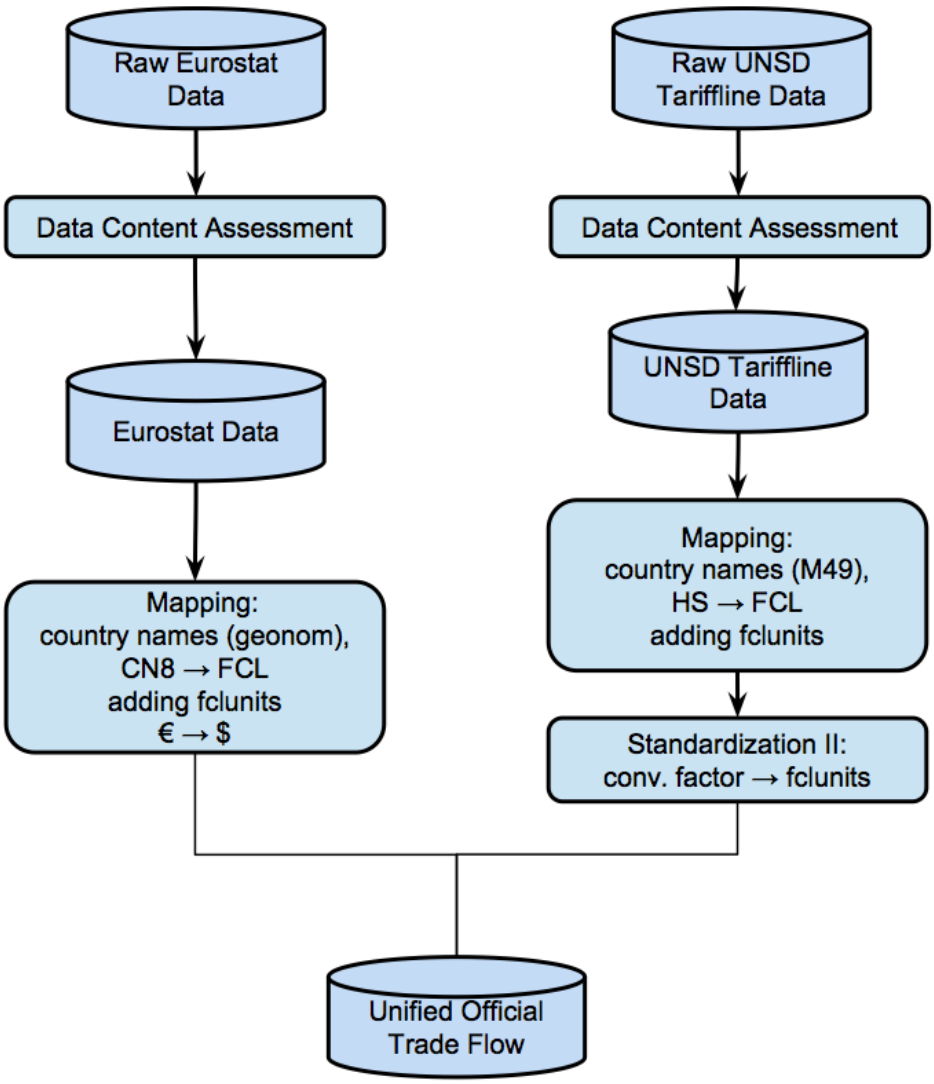
- Conversion of FCL into CPC codes. This conversion is based on the table of conversion 2.1 expanded. If some FCL codes are not mapped into CPC ones, the corresponding records are filtered out.
- Concersion from FAO country code to M49. As before, if some countries are not correctly mapped, they are filtered out from the final output.
- The results are gathered based on elements codes available on the SWS. Before this step each record presents both quantity and value, while after this step each record represents or a quantity or a value.

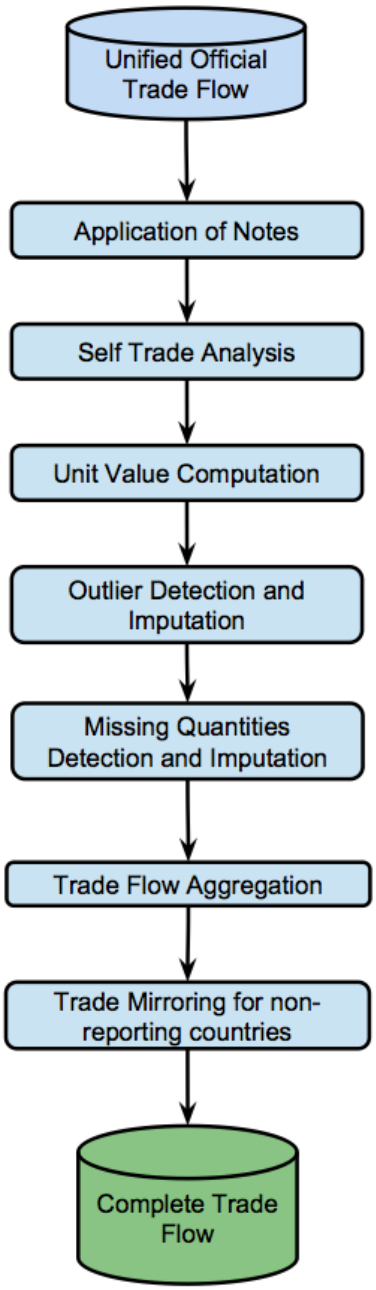
The first submodule save the final output in the `completed_tf_cpc_m49` dataset, within the `trade` domain.

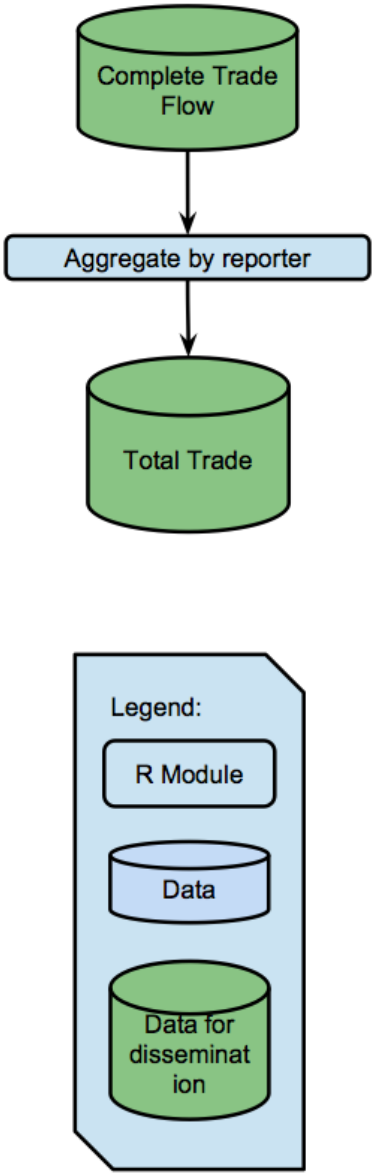
2. Total trade CPC

This second submodule uses as input the output of the previous submodule. These two modules are separated because the two outputs are needed for different scopes.
This module aggregates by reporters over partners countries to a single total trade for each unique CPC commodity code.
The module save the ouput into the dataset `total_trade_cpc_m49`, within the `trade` domain.

3. Flow Chart Process




Unified Official Trade Flow
Application of Notes
Self Trade Analysis
Unit Value Computation
Outlier Detection and Imputation
Missing Quantities Detection and Imputation
Trade Flow Aggregation
Trade Mirroring for non-reporting countries
Complete Trade Flow


Complete Trade Flow
Aggregate by reporter
Total Trade
Legend:
R Module
Data
Data for disseminat ion

4. Future work

In this section all open problems and future work is discussed.
To the reader who is supposed to code: please refer to the GitHub Issue page as more technical open issue.

4.1. Validation Steps

This section represents the most high priority task for the trade module.

Raw Data - Data content assessment

The pre-analyses of the assessment of the data has to be integrated in the module. In the vignettes folder, a sample of pre-analysis is given, but not integrated in the main module (file name: `preanalysis_2009.Rmd`). The pre-analysis script calculate the total number of recordsfor both Eurostat and UNSD Tariffline datasets and the distribution of lenght of the commodty HS codes (for UNSD Tariffline) and CN8 for Eurostat is performed. For each country we report if data includes imports, exports, re-exports and re-imports at all possible lenght.
All records with hs-lenght (for UNSD Tariffline) or CN8-lenght (for Eurostat) less than 6 are removed.
The pre-analysis script produces a html file that can be read in any browswer, but a suitable solution for the SWS is needed.
This script should be part of the validation steps within the module.

Report and check of discarded elements after mapping

Each mapping routine might produce some unsolved mapping. At the moment the module is saving the unsolved mapping records in a separated variable, but not reported. All unsolved mapping should be reported and possibly solved in the future.

Destination Table

The `complete_tf_cpc` module produces output for all the records passing all the routines and not filtered out. The module does not check if any commodity is missing. A possible solution would be to have a destination table with all the commodities of interest and the module should fill the destination table. In this way the output validation step should be achieved.

4.2. Unbalanced World Trade Matrix

Before the aggregation, a trade imbalances report might be produced.

4.3. Time Series Analysis

Upon availability of time-series data, a check of the CPC-based unit values accross the time series should be performed. Differentiation between errors in the order of magnitude (time series) and outliers (cross section). An additional submodule of imputation of missing data using time series analysis would be a solution.

4.4. CIF/FOB

The CIF/FOB correction for mirroring is, at the time being, set up to 12%. This has been

suggested by team B/C, but additional work might be done in order to assess if the estimate is appropriate. There might be different range of percentages for different type of countries and by distance between reporters and partners.
A study can be conducted on available records on both side: this means records for which the commodity is reported by the reporter and by the partner.

4.5. Re-import and Re-export

At the moment all re-import and re-export is considered as, respectively, import and export. More study might be conducted in order to identify countries more prone to report re-import and re-export.

4.6. Self Trade Analysis

For all the records having the same reporter and partner an analysis is performed on a separate script within the `vignette` folder: `selftrade.R`. This might be incorporated in the module and might produce suitable output within the SWS.
The sum of the value is computed for both countries and commodity, in order to spot out the countries reporting massive self trade and which are the main commodity reported as self trade. Summary statistics are computed world wide.

4.7. Pseudo-automatic mapping of commodities

An additional method has to be added in the future: the algorithm should try to trim the code not mapped and try to map them with shorter HS codes. If any of shorter codes (from right to left) are then not mapped, we can definitely discard the record. If a specific record has a HS code not mapping to any specific FCL code, then the record is reported and removed.

4.8. Mapping from HS to FCL

In the module for commodities we have 2 different mapping. From HS to FCL, using mapping produced by team B/C and then from FCL to CPC 2.1. This mapping is available for the following years: 1997, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013. The following years are missing: 1998, 1999, 2000. The mapping for 2014 has been copied from year 2013, but the results need to be checked.
In the future direct mapping from HS to CPC has been asked from managment. A possible solution with the mapping to CPC has been sent to Carola (09.06.2016), but anyway this needs revision.

4.9. Mapping from Comtrade M49 and Geonomenclature directly to M49

The country codes, as the commodity ones, have two step of mapping. This results in higher risk of data loss due to unsolved mapping. In the future a direct map from Comtrade M49 (Tariffline UNSD) to M49 and from Geonomenclature (Eurostat) to M49 would be ideal.

4.10. Flag correction

Flags: when mirroring is performed, the quantity will stay official, while the value will change flag (high priority)

4.11. Outlier coefficient

The outlier coefficient is set up to 1.5. The outlier coefficient is a input parameter of the

`complete_tf_cpc` submodule. After discussion with team B/C (23.06.2016) a specific analysis has to be perfomed to understand what is the best coefficient to be used in order to reflect old results. After this analysis, the outlier coefficient should be hard-coded within the code of the module without letting the user to modify it anymore.

4.12. Food-aid

This has to be incorporated also to understand the trend in a time series analysis. This needs special study to understand if we can get the data just from the exports not reported as imports in the partner.

Disclaimer

This Working Paper should not be reported as representing the official view of the FAO. The views expressed in this Working Paper are those of the author and do not necessarily represent those of the FAO or FAO policy. Working Papers describe research in progress by the authors and are published to elicit comments and to further discussion.

This paper is dynamically generated on July 27, 2016 and is subject to changes and updates.

Affiliation:
Marco Garieri
Economics and Social Statistics Division (ESS)
Economic and Social Development Department (ES)
Food and Agriculture Organization of the United Nations (FAO)
Viale delle Terme di Caracalla 00153 Rome, Italy
E-mail: marco.garieri@fao.org
URL: https://github.com/SWS-Methodology/faoswsTrade