faoswsTrade: Data Sources

Marco Garieri
Food and Agriculture Organization
of the United Nations

Abstract

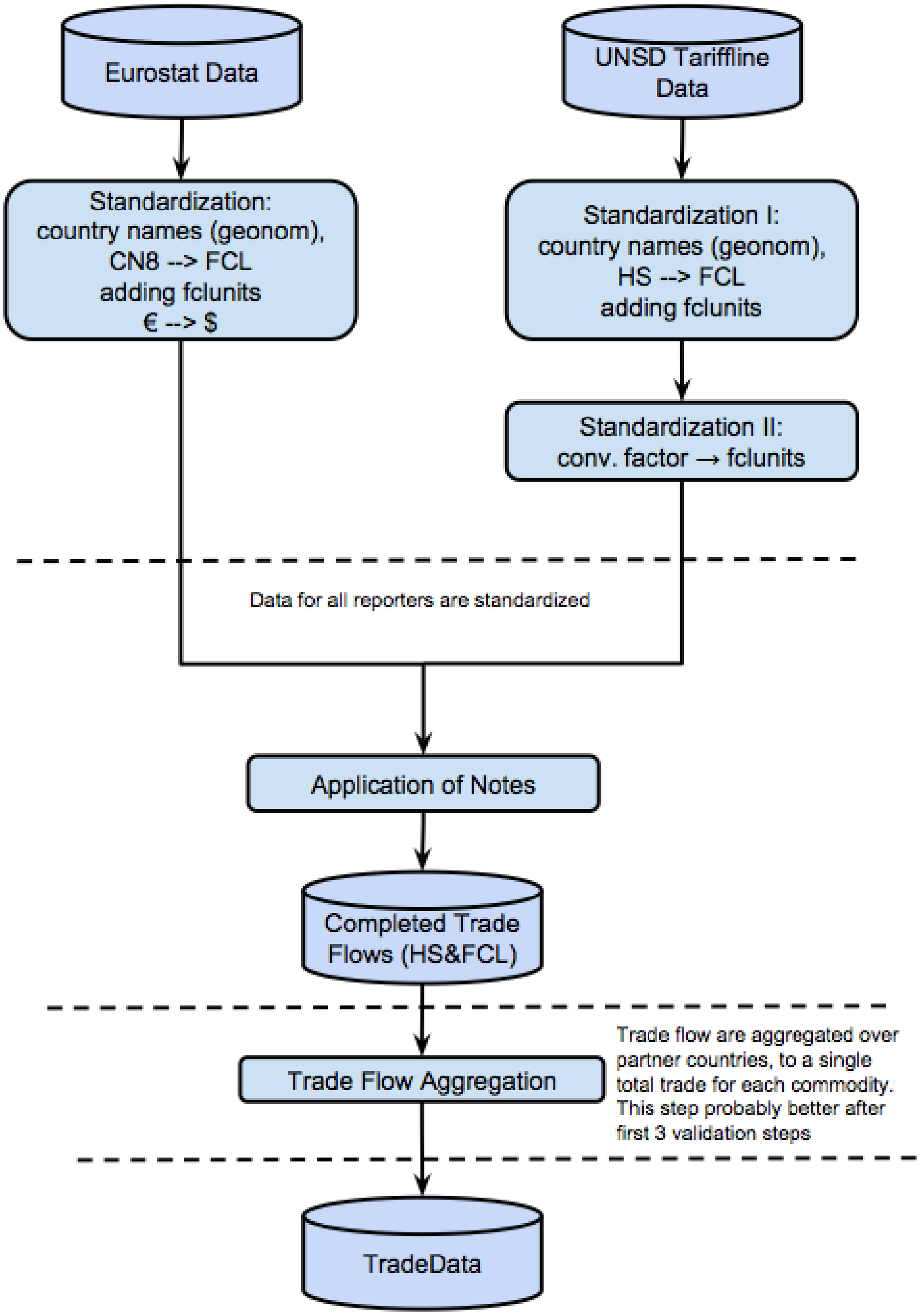
This vignette provides a detailed description of the various data sources used in the trade modules.

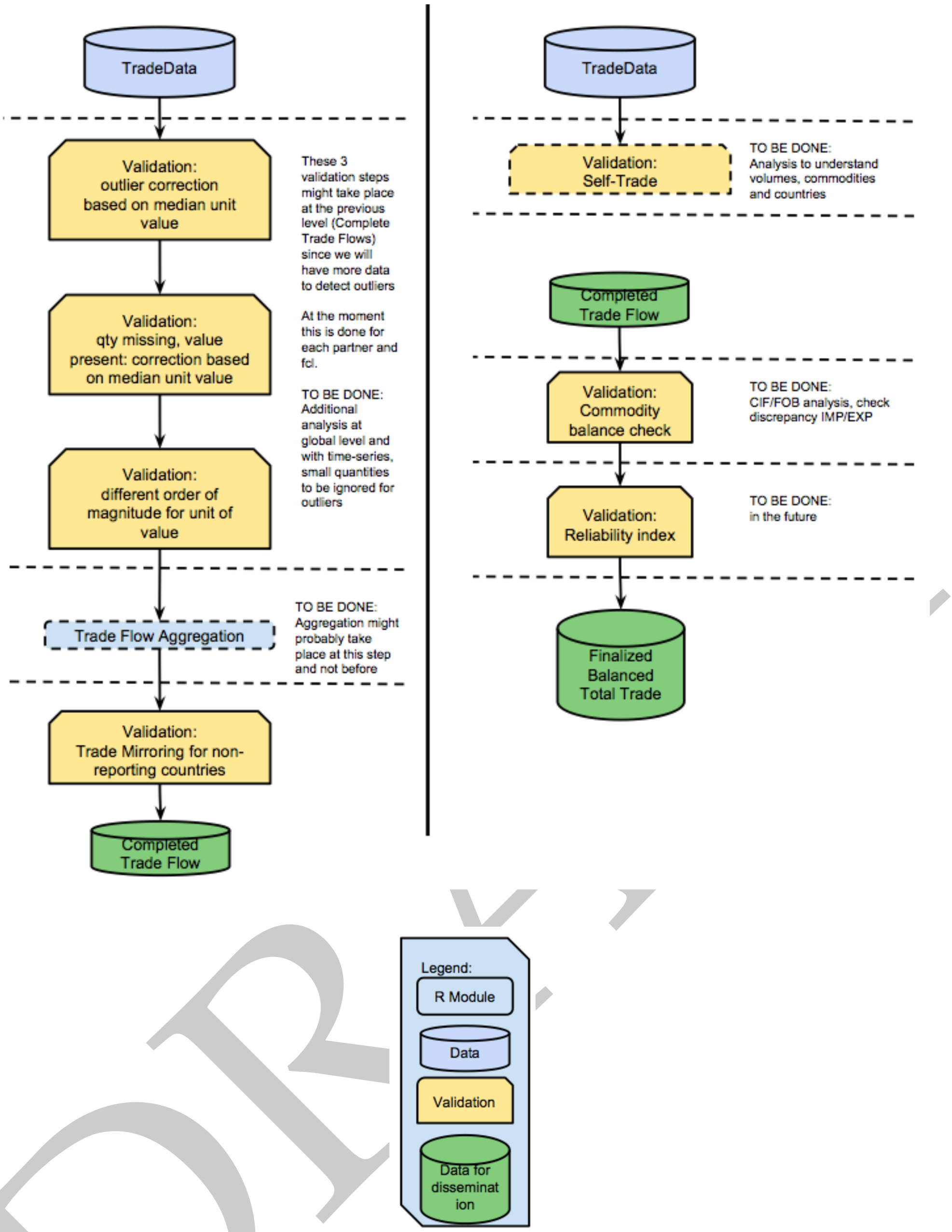
Keywords: Agricultural Trade, Tariff Line, Eurostat, Mirroring.

1. Data Source

1.1. Flow Chart

Description of the entire flow


Eurostat Data
UNSD Tariffline Data
Standardization: country names (geonom), CN8 --> FCL adding fclunits € --> $
Standardization I: country names (geonom), HS --> FCL adding fclunits
Standardization II: conv. factor → fclunits
Data for all reporters are standardized
Application of Notes
Completed Trade Flows (HS&FCL)
Trade Flow Aggregation
Trade flow are aggregated over partner countries, to a single total trade for each commodity. This step probably better after first 3 validation steps
TradeData


TradeData
Validation: outlier correction based on median unit value
These 3 validation steps might take place at the previous level (Complete Trade Flows) since we will have more data to detect outliers
Validation: qty missing, value present: correction based on median unit value
At the moment this is done for each partner and fcl.
TO BE DONE: Additional analysis at global level and with time-series, small quantities to be ignored for outliers
Validation: different order of magnitude for unit of value
Trade Flow Aggregation
TO BE DONE: Aggregation might probably take place at this step and not before
Validation: Trade Mirroring for non-reporting countries
Completed Trade Flow
TradeData
Validation: Self-Trade
TO BE DONE: Analysis to understand volumes, commodities and countries
Completed Trade Flow
Validation: Commodity balance check
TO BE DONE: CIF/FOB analysis, check discrepancy IMP/EXP
Validation: Reliability index
TO BE DONE: in the future
Finalized Balanced Total Trade
Legend:
R Module
Data
Validation
Data for disseminat ion

1.2. Data

Data received from Giorgio on March 18 for UNSD and Eurostat from 2009 to 2014. Some prefiltering has been done from Giorgio:

Eurostat – code of reporter (declarant) just numeric (letters are not allowed)
– code of partner (partner) just numeric (letters are not allowed)
– code of hs (product_nc) just numeric (letters are not allowed)

UNSD – code of hs (comm) just numeric (letters are not allowed)

This is the summary of the data received:

Eurostat 2009 Number of records: 8665414, distribution of digits: 2 - 446403 (5%), 8 - 8219011 (95%)

2010 Number of records: 8861460, distribution of digits: 2 - 449191 (5%), 8 - 8412269 (95%)

2011 Number of records: 8977093, distribution of digits: 2 - 458032 (5%), 8 - 8519061 (95%)

2012 Number of records: 9299996, distribution of digits: 2 - 468915 (5%), 8 - 8831081 (95%)

2013 Number of records: 9493511, distribution of digits: 2 - 476782 (5%), 8 - 9016729 (95%)

2014 Number of records: 9610127, distribution of digits: 2 - 480525 (5%), 8 - 9129602 (95%)

UNSD 2009 Number of records: 40262359, distribution of digits: 2 - 62687 (0.2%), 4 - 822372 (0.02%), 5 - 63 (0%), 6 - 7197277 (18%), 7 - 331 (0%), 8 - 20218041 (50%), 9 - 722018 (2%), 10 - 6789612 (17%), 11 - 2769625 (7%), 12 - 1680333 (4%)

2010 Number of records: 46654452, distribution of digits: 2 - 30547 (0.1%), 4 - 199688 (0.4%), 5 - 0 (0%), 6 - 12394503 (27%), 7 - 54685 (0.1%), 8 - 21595733 (46%), 9 - 608860 (1%), 10 - 9102501 (20%), 11 - 2667935 (6%), 12 - 0 (0%)

2011 Number of records: 63535135, distribution of digits: 2 - 41486 (0.1%), 4 - 1640555 (2.6%), 5 - 1 (0%), 6 - 14427100 (23%), 7 - 8447 (0%), 8 - 28948926 (46%), 9 - 636035 (1%), 10 - 13379759 (21%), 11 - 2645793 (4%), 12 - 1807033 (3%)

2012 Number of records: 66175819, distribution of digits: 2 - 42165 (0.1%), 4 - 131569 (0.2%), 5 - 0 (0%), 6 - 18723116 (28%), 7- 14986 (0%), 8 - 32048866 (48%), 9 - 643335 (1%), 10 - 13565552 (21%), 11 - 1006230 (1.5%), 12 - 0 (0%)

2013 Number of records: 70075550, distribution of digits: 2 - 25395 (0%), 4 - 995 (0%), 5 - 0 (0%), 6 - 26224495 (37%), 7 - 16979 (0%), 8 - 32652742 (47%), 9 - 654930 (1%), 10 - 9765518 (14%), 11 - 734496 (1%), 12 - 0 (0%)

2014 Number of records: 79728175, distribution of digits: 2 - 59500 (0.1%), 4 - 222713 (0.3%), 5 - 0 (0%), 6 - 25279829 (32%), 7 - 66423 (0.1%), 8 - 42679600 (54%), 9 - 649753 (1%), 10 - 10088068 (13%), 11 - 682289 (1%), 12 - 0 (0%)

1.3. Example of tables

1.4. Process

1) Raw UNSD Tariffline Data

This section covers pre-processing operations. Strictly speaking, pre-processing operations do not pertain to the trade module but to input data management, editing and cleaning. We fully agree that more clarity is needed on the operations performed, on the workflow and on who is responsible for them. In principle, these operations should be carried out who manages data import in the SWS. At present, these operations have been performed partly by the SWS team, partly by the developer. The pre-processing operations will be included in a sub-routine. The data content assessment will be done systematically as a summary table and an automatic report.

2) UNSD Assessed Tariffline Data and Eurostat Data

a. describes well the operations undertaken by the module. It must be clear that no aggregations are done at this stage, i.e. no information is lost. Standardization and mapping steps create additional columns. b. There will be no aggregation at this stage. The large number of records strengthens the outliers detection procedure. c. The module already generates a report on missing links. What needs to be discussed and agreed with Team B/C and the classification experts is how to up-date the correspondence tables, i.e. the maps. d. i. Capturing missing links. Work had already started in this direction. A sub-module in the trade module tries to map HS codes with FCL automatically. It first checks if a 0 values is missing on the left of the code, then is looks at the highest levels of the classification to do the mapping. It must be said that the most problematic items for this sub-modules come from fisheries, pesticides, herbicides and fertilizers. While it is not particularly affecting FBS data, the matter must be solved. Point d. iii. Very good suggestion Some preliminary analysis were already made in this direction. The application of natural language processing takes time because it requires HS metadata, i.e. downloading HS labels on top of HS codes. Labels are not included in the file download for size reasons and speed. On can consider a separate extraction of single HS codes and descriptions and develop a sub-module that works on this file instead of the large data file. Its implementation cannot a priority now. e. The comment made describes how the module works. 3) Unified Official Trade Flows Dataset

This step is called âĂIJComplete Trade FlowsâĂİ in the flowchart and should be called so unanimously to avoid confusion. b. Validation steps i. Changing order between correcting for Orders of Magnitude and Detecting Outliers is a very valid suggestion. The following changes will be made to the module. The suggested process for Order of Magnitude corrections, however, is not viable because will excessively slow down the module. The following change will be made to the module: corrections for Orders of Magnitude will be managed by the outlier detection process. The process will proceed in steps. 1. Orders of magnitude will be detected through time series analysis. The sub-module is still being developed. 2. When detecting an outlier in cross-section data, the module will search first for mirror transactions. If there exists a transaction between the same trading partners (same commodity but opposite flow) whose quantity matches but for a multiple of 10, then the module will correct the outlier by copying the quantity recorded by the trading partner. 3. If there are no mirror quantities, the module will apply the standard outlier correction process.

iii. Validation: Outlier Correction The module is already implementing a similar threshold rule, to keep as much official data as possible. The test however is implemented at tariffline level and not at hs level. This way, commodities are homogenous and extreme unit values are more likely due to pure price effects than to product characteristics. c. Output HS trade table The module produces this intermediate output but does not print it for the sake of efficiency and space saving. To be discussed.

e. There should be no automatic over-writing at FCL level, as suggested, once the unit values have already been checked at corrected at tariffline level.

f. The module will have a sub-module to measure CIF/FOB differences and check the 12

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://gitlab.com/faoess/tradeproc