Appendix: `trade_validation_cpc` module

Christian A. Mongeau Ospina
Food and Agriculture Organization of the United Nations

9 March 2018

This document shows the operations that the `trade_validation_cpc` module carries out in order to build the dataset used in the validation of trade flows.

Parameters

- `startyear` (SWS label: "Start year"): first year for processing.
- `endyear` (SWS label: "End year"): last year for processing.
- `useprevious` (SWS label: "Use previous data?"): if set to `TRUE`, previously downloaded data will be used, otherwise a new query will be performed. *This should always be `FALSE`, leaving `TRUE` as a viable option only for testing/debugging purposes.*
- `multicore`: if set to `TRUE`, operations will be performed with parallel programming, i.e., sequential operations will be distributed to the different cores of the machine on which the module runs. This implies a reduction of computation time. *This is a hardcoded parameter.*
- `threshold`: threshold used to define what an outlier is (see below). *This is a hardcoded parameter.*
- `morder`: order of the moving average of unit values. *This is a hardcoded parameter.*

Set up parallel processing requirements

If `multicore` = `TRUE` some preparatory stepd need to be performed: Initialise the required packages and export objects to the cores.

Download data

A query to the SWS with all reporters, partners, items, elemtents is performed, by reporter. The result of each reporter-specific query gets completed for all years and for all existing partner/flow/item combinations. The expansion for all years is necessary as calculations that need a complete time series (even with NA values) is required.

Once a reporter-specific dataset is downloaded and completed, moving averages of unit value and the ratios of the unit value with respect to these averages are calculated. A first version of "outlier" (*outn*) that uses information only at the reporter-level is calculated as:

$$outn = \begin{cases} \text{TRUE}, & \text{if } ratio < 1 - threshold \text{ or } ratio > 1 + threshold \\ \text{FALSE}, & \text{otherwise} \end{cases}$$

where *threshold* is the `threshold` parameter (as of 2018-03-09 it is set to 0.5), $ratio = uv/\overline{uv}$ ($\overline{uv}$ is the `morder` (3 as of 2018-03-09) moving average of uv).

The reporter-specific datasets get stored in the "shared drive" of the server for latter use.

Data assembly and further operations

Once all reporter-specific datasets are available, they are assembled together in order to compute statistics that uses all reporters' information, specifically, the following outliers are identified:

- `outmw100`: equal to `TRUE` if the unit value is less or greater than the median unit value of the item for all reporters;
- `outM`: a boxplot approach is used to define this outlier. In particular, if the unit value falls outside the boxplot whiskers it is considered an outlier. To overcome some asymmetry in the the distribution of unit values, they are transformed into logarithmns.
- `outp`: defined as *outn*, but the threshold is not fixed for all items, but is item-specific and the lower and upper thresholds correspond to the 5th and 95th percentile of the distribution of the unit value for the item for all reporters and flows.

By default, the "outlier" variable is defined to be equal to *outn*, though in the validation tool it is possible to choose which method to use when defining what an outlier is.

Save data

Data is ready to be saved. Before doing so, some informational variables are added:

- `perc.value` and `perc.qty`: give the importance (percentage) of values and quantities of a specific item in the reporters' total trade.
- reporter, partner, and commodity names: SWS does not return the names of the dimensions, so this information is retrieved and joined to the dataset.

After the additional information is added, the dataset is saved to the "shared drive" from where the validation tool will read it.