**Interactive tools for trade data validation**

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

15 September 2017

The interactive tool for trade data validation consists in two modules:

* [Bilateral trade flow](http://hqlprsws1.hq.un.fao.org:3838/outliers/)
* [Module/FAOSTAT comparison, flows](http://hqlprsws1.hq.un.fao.org:3838/flows/)

As the names suggest, the first one deals with bilateral trade flows (i.e., at the "complete trade flow" level), while the second deals with comparisons of module and FAOSTAT aggregated data (i.e., at the "total trade flow" level).

By using the tools (singularly, or, possibly, in combination one with each other) trade data can be scrutinised and eventually corrected.

In the next sections, the modules will be explained in detail.

While indeed the bilateral and total trade tools for validation lives in two different pages, the latter can be accessed by the former in the "Module/FAOSTAT flows" component (technically, this is an *iframe* that opens the "flows" app inside the bilateral validation tool).

**Note: given that the process for generating the whole elements of *Module/FAOSTAT comparison* is very time-consuming, currently other validation strategies for total trade are in course of development.**

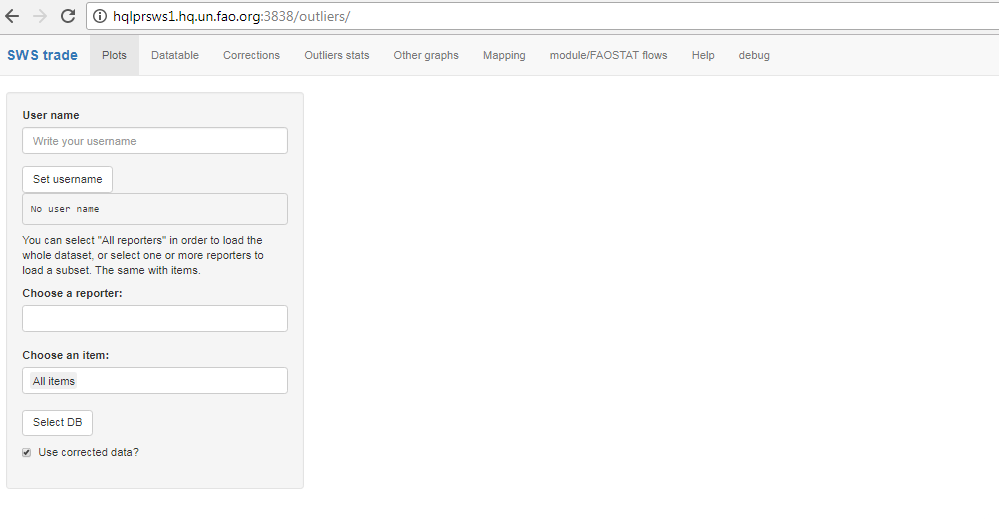
## Bilateral trade flows

The interactive tool for validating bilateral trade flows is available at:

<http://hqlprsws1.hq.un.fao.org:3838/outliers/>

The first time the tool is accessed, the complete dataset with bilateral trade flows will be loaded. Given that it consists of more than 14 million of bilateral transactions, it will require some seconds to load (around 30 seconds). Once it is accessed by someone, the whole dataset remains in the server cache until there is at least one analyst using it.

When you load the app, it looks like the image below.



Initial state

In order to access the functionalities of the tool, you will have to write your user name in the "User name" field, then select a reporter (or more reporters) and and item (or "All items", which is the default). When you do this, you can click on "Select DB" so that your choice is saved.

The app is composed by seven components:

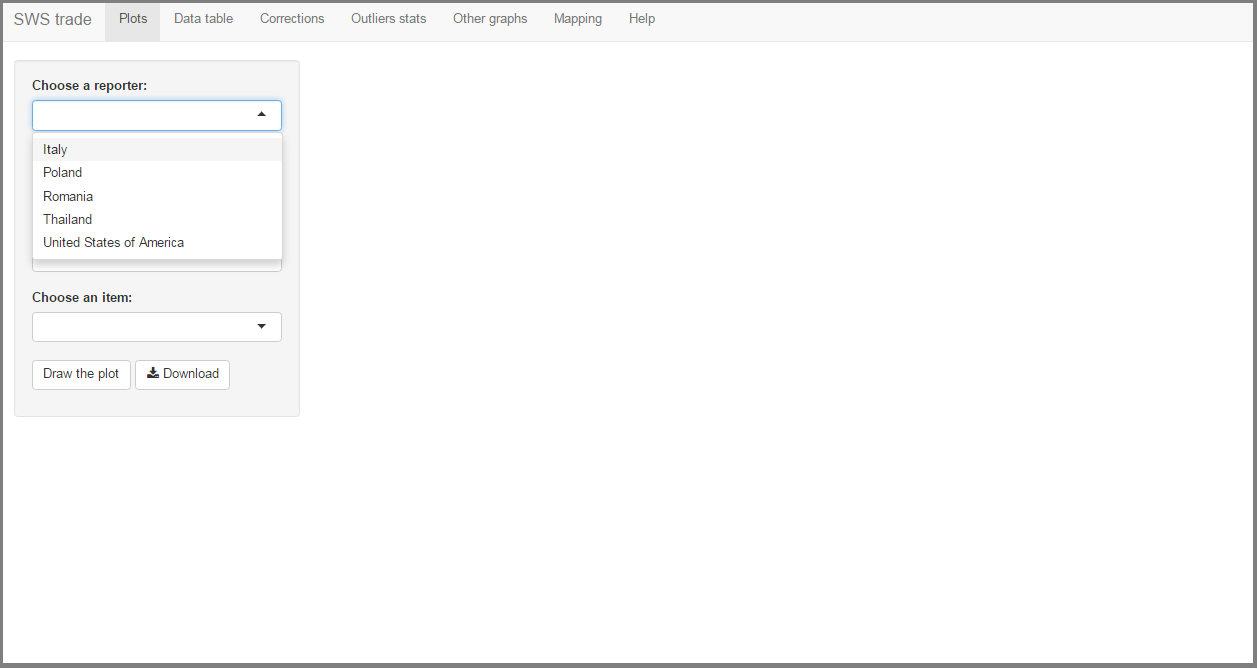
* **Plots**
* **Datatable**
* **Corrections**
* **Outliers stats**
* **Other graphs**
* **Mapping**
* **Module/FAOSTAT flows**

The next sub-sections explain the various components.

### Plots

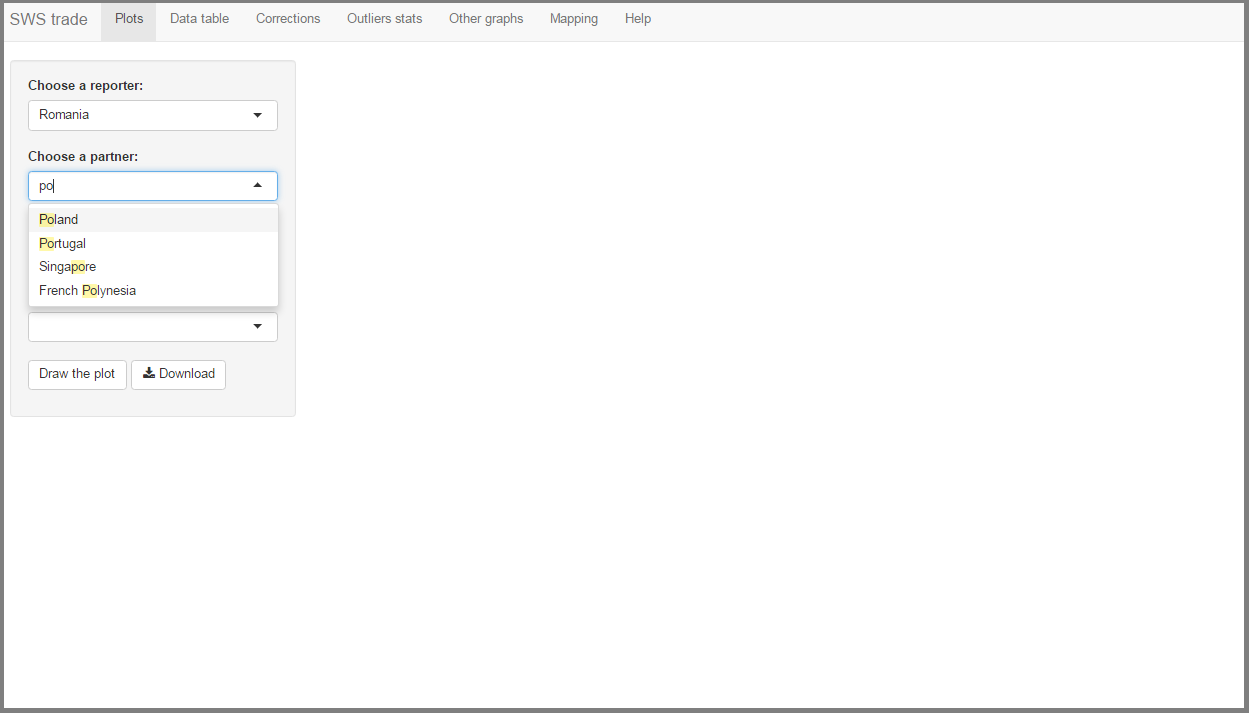
In this component you can plot graphs of the unit value, quantity and value for a combination of "reporter", "partner", "flow" (1 = imports, 2 = exports) and "item", which can be selected in the menu on the left column or, more likely, in the "Datatable" table where the outliers are shown (you'll have to click on "link"; details further in the text).

When you click on the menu under "Choose a reporter:", a list of reporters appears and you can select one reporter.



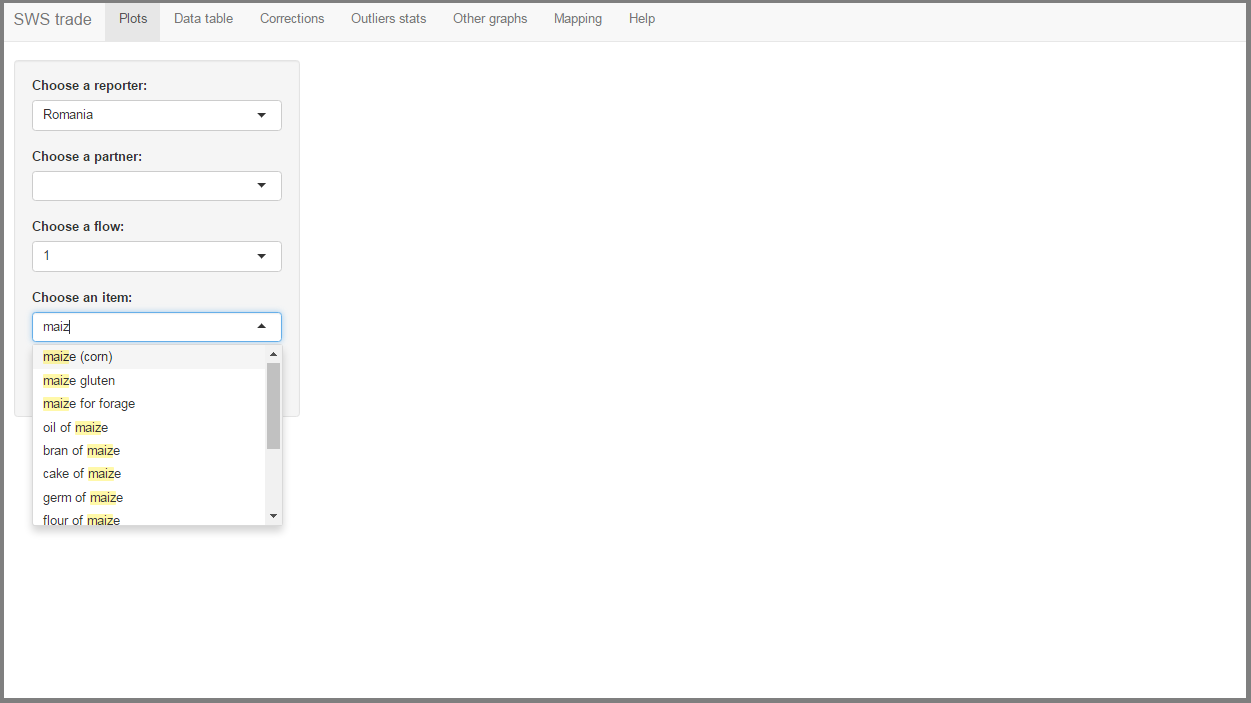
Reporter selection

The same for partners (notice that you can write the name of the country and those that match will remain visible).



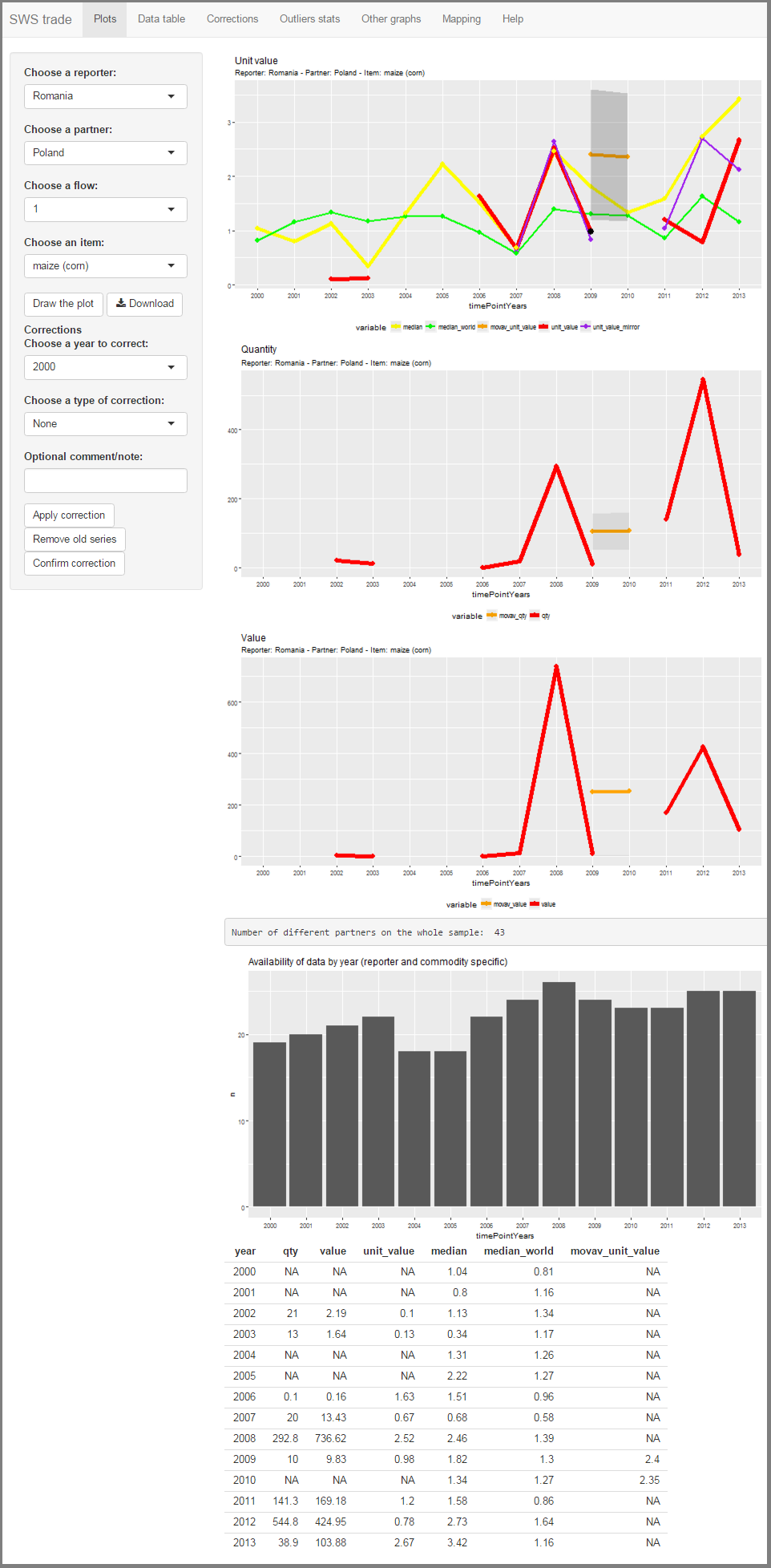
Partner selection

A similar approach should be used for selection the item (commodity).



Item selection

Once you define your selection, you can click on "Draw the plot" to see the plots (note that you can also choose a combination of reporter/partner/flow/item in the "Datatable" component; see below). An example of the results can be seen in the following graph.



Results

In the first three graphs, colors represent (not all variables are shown in all graphs):

* **red**: original variables (i.e., unit value, quantity, and value)
* **orange** moving averages.
* **purple**: mirror flow (if/when it exists).
* **black**: "corrected" series (will show up once a correction is done; see next section).
* **yellow**: median with respect to all reporters.
* **green**: median of the world.

Outliers, defined by using the "fixed threshold" method (see the definition below in the "Datatable" paragraph), are shown with a black point.

Below the first three graphs, there is an infobox that shows the number of countries with which the reporter has had at least one flow of that commodity in the whole period condidered (in the example it says: "Number of different partners on the whole sample: 43"). Just below this information, there is a graph showing the distribution of actual partners by specific years.

There is also a graph showing the aggregated flow of the selected commodity for the repoter under analysis.

Finally, a snapshot of the data is shown at the end of the page.

After you draw the plots, it is possible to download the underlying data by clicking the "Download" button.



Download data

Moreover, in this screen you can do corrections to outlying observations.

Note: instead of selecting combinations of reporters/partners/items in order to search for outliers, it is possible to select the series with outliers in the "Datatable" component, which will be described further. This allows to save some time by focussing only on countries with outliers. You may want to have a look at that section before reading how to actually correct outliers, as it shows how the outliers are identified.

#### Choose a correction type

If you need to correct some anomalous (outlier) observation, you can do it in the "Plots" component.

It is possible to correct either the value or the quantity. In both cases, the main value that gets corrected is the unit value to and, consequently, the variable you decided to correct by considering the other variable as already correct. Thus, if you decide to correct a quantity, you first should have a look at the unit value of that transaction, fix it and then the imputed quantity will be given by the ratio between the value and the corrected unit value; conversely, if you want to correct a value, you fix the unit value and the imputed value is obtained by multiplying the quantity by the corrected unit value.

You can choose the year for which you want to do a correction ("Choose a year to correct:") and a type of correction ("Choose a type of correction:"):

* **None**: This should be used to confirm that an outlier is OK and should not be corrected.
* **Measurement factor**: This can be useful if the quantity or value may have been reported with a wrong measurement unit, by a factor of 10. If you choose this option, it will be displayed below if the ratio of the unit value to the median unit value is "near" a factor of 10 and will suggest which factor is the most likely. More technically, the ratio between the unit value and the median unit value is computed; if this ratio is greater than or less than (with a 10% approximation) then it means that the transaction was likely reported with a unit of measurement (of a factor of 10) issue.
* **Mirror flow**: Use the flow of the partner, if it exists (if it does not, you will get a feedback indicating that).
* **Outlier correction**: It is possible to use three different types of correction (in all cases, the values of these figures will be printed, so that you can choose the most appropriate):
  + **Median partners**: median with respect to all partners, that is, the median for the reporter's trade (import or export specific) in the year of interest of the analysed commodity to all its partner is used.
  + **Median world**: median of the unit value of all reporters to all partners, that is, the median of all the reporters that traded (import or export specific) the analysed commodity in the year of interest is used.
  + **Moving average**: the three year-moving average (starting from the year previous to the one analysed) of the unit value of the trade flow of the commodity of the analysed reporter.
* **Expert knowledge**: literature or otherwise justified figure, in other words, this is a figure that an analyst set given some external source.

In the image below, the year was set to 2012, and the menu for the correction type shows the alternatives.



Correction menus

In "Optional comment/note:" you **can** write a comment that will be saved into the corrections table. If the type of correction is set to "Expert knowledge" then the note becomes "Mandatory comment/note" and the comment **must** be written (basically, if an expert gave a specific number, the name of the expert and/or the source of the data must be written here).

You can apply the correction by clicking "Apply correction". In this case the unit value and quantity (or value) plots will show a new series which contains the correction.

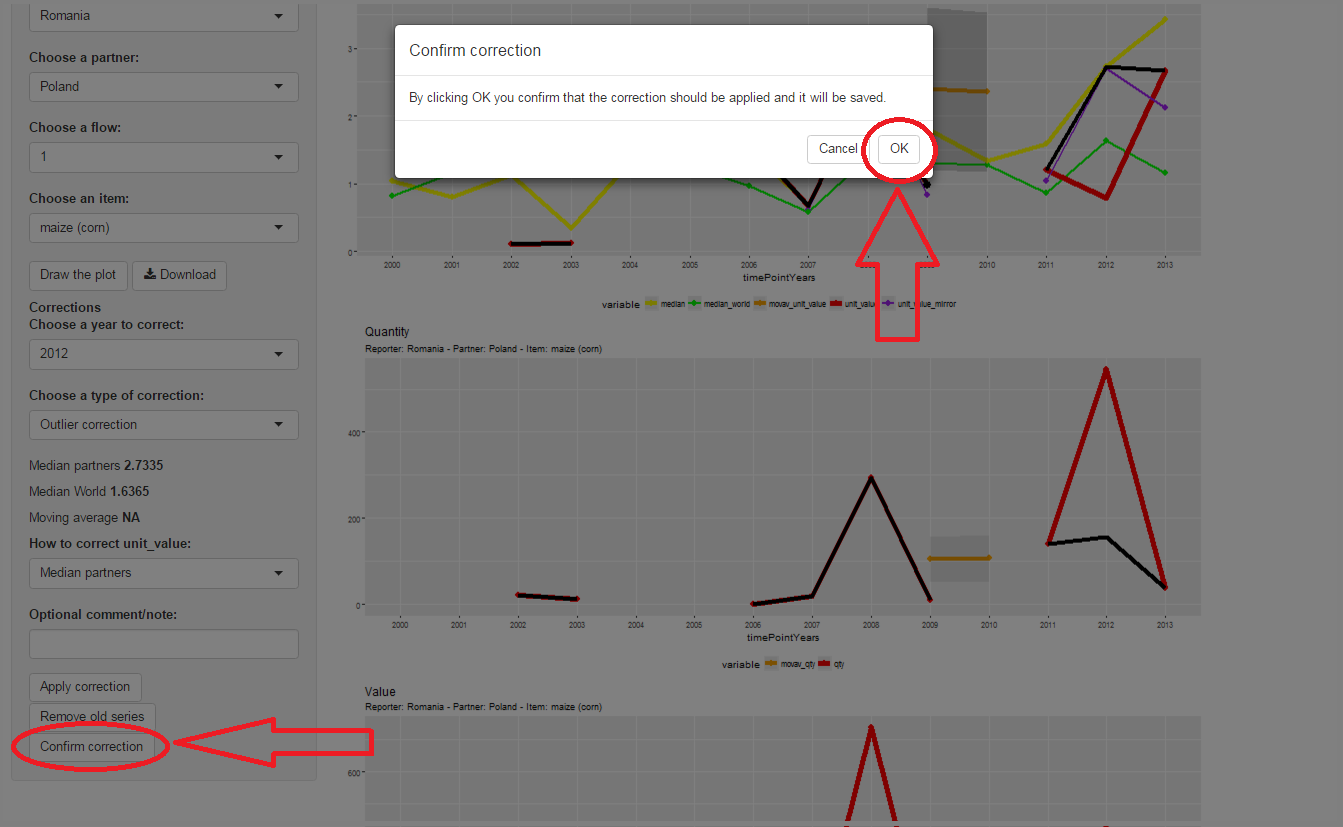
The following figure shows a correction example: in 2012 we suspect that the quantity variable is too high, which is reflected in a low unit value. We set the new unit value to be equal to the median unit value with respect to all partners (Choose a year to correct: > 2012; Choose a type of correction: > Outlier correction; How to correct unit\_value: > Median partners). A black line will show the new value (the original value is still visible, in red).



Corrrection example

You can remove the old (or original) series by clicking the "Remove old series" button. If you want it back, click again on the same button.

Once you are confident with the correction, click "Confirm correction". This will open a confirmation menu, where you can confirm (OK), or go back (if you decide to go back, no correction will be stored . In the case you confirm, the correction will be saved in the correction table that can be visualised in the "Corrections" component (see below).



Confirm correction

### Datatable

This table presents all the observations classified as outliers. In the current configuration of the app, the definition criterion for which an observation is considered as an outlier can be chosen by the user (when you choose a criterion here, this will be used also in the "Outliers stats" and "Other graphs" component, not in the "Plots" component). We propose four criteria:

* **Fixed threshold**: (Default method) Observations are outliers if the ratio of the unit value to its 3-year moving average is below 0.5 or above 1.5 (the three year period starts in the year immediately before the analysed year; the moving average for the first three years is computed as a "forward" moving average).
* **Variable threshold**: Similar to the "Fixed threshold" method (see above), but thresholds are defined as the 5th and 95th percentiles of the distribution of the ratio computed specifically for the item (there are different thresholds, e.g., for "maize (corn)", "mushrooms and thruffles", etc.). In other words, for each commodity the distribution of the unit value and the three-year moving average is obtained and the thresholds are set as the 5th and 95th percentile.
* **100 median**: A unit value is considered an outlier if it is below 1/100 or above 100 times the median unit value of all reporter/partners (i.e., flow/item/year specific).
* **Boxplot**: A unit value is an outlier if its logarithm is outside the intervals defined as and , where Q1 and Q3 are the first and third quartile of the logarithm of the unit values of that specific item/flow combination with respect to all partners of the reporter.

If, sor some reason, you do not want this table to show only the outliers, but *all* the dataset (i.e., both outlier and non-outlier observations) then it is possible to select the **All data** in the dropdown menu under "Choose an outlier detection method:".

Each row has a "link" in the leftmost column: if you click this, the reporter/partner/flow/item combination will be shown in the "Plots" tab.

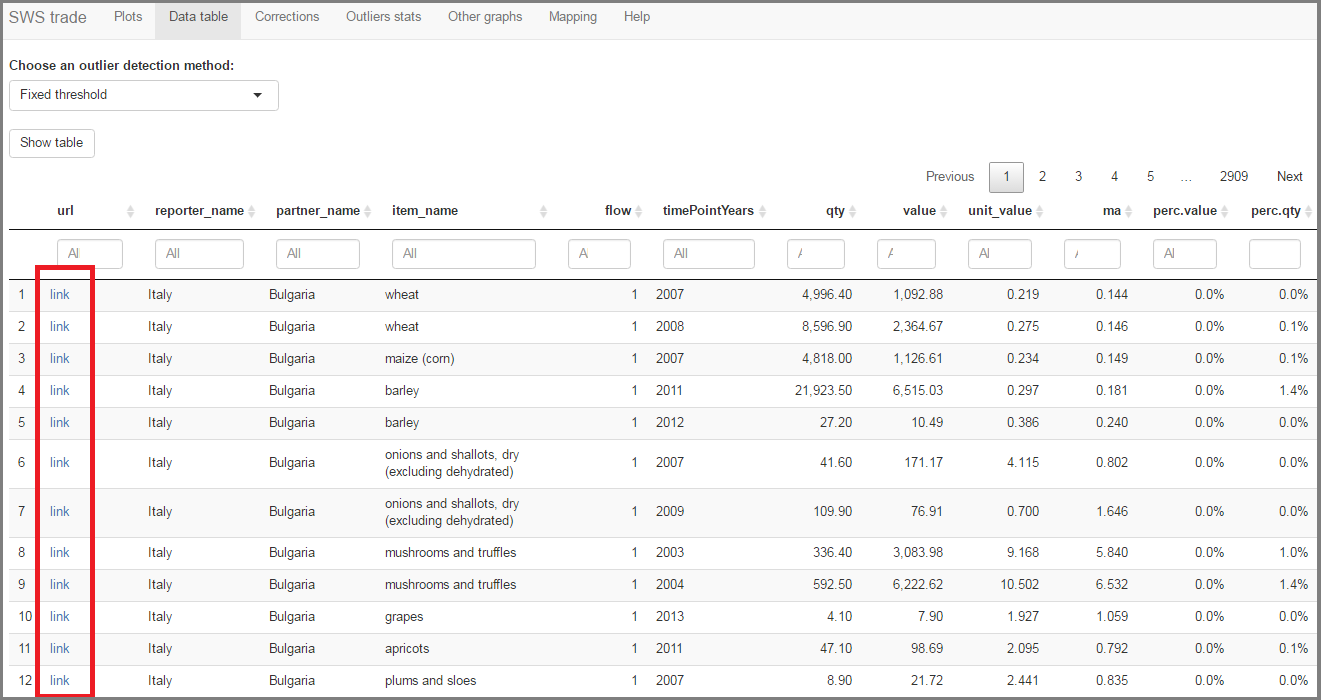


Table with outliers

The meaning of the columns is self explanatory, except:

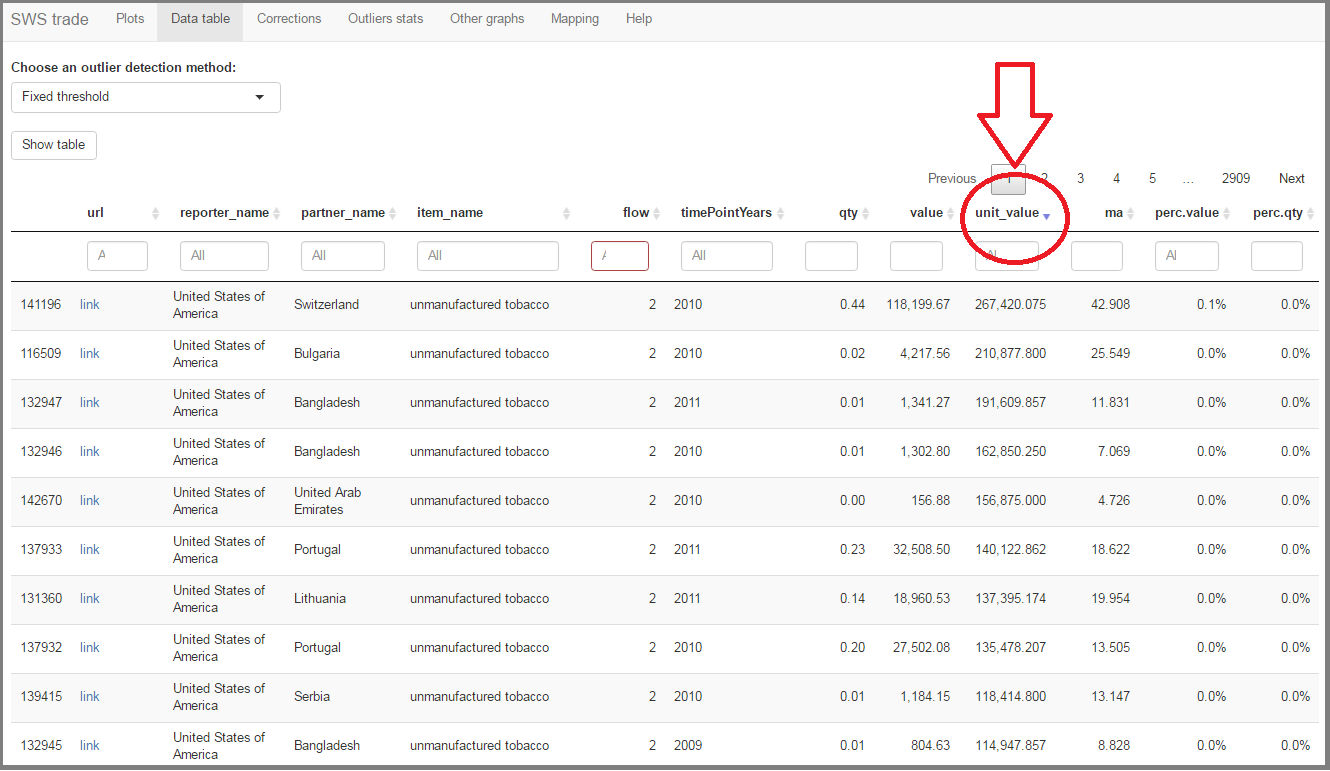
* **ma**: The 3-year moving average of the unit value.
* **perc.value**: Is the percentage of the value of the transaction over the total value of the flow for the reporter (e.g., the value of US imports of maize from Mexico over the total value of the imports of the US).
* **perc.qty**: Percentage of the quantity of the transaction over the total quantities traded by the reporter for that item (e.g., the quantity of US imports of maize from Mexico over the total quantity of maize imported by the US).
* **corrected**: shows whether the observation has been already corrected.

The table can be filtered by using the table headings. For instance, let us say that only the imports of the US from Mexico requires analysis: in this case, "United States of America" (or just some part of the whole name) should be written in the field below "reporter\_name", "Mexico" (or a part of the name) below "partner\_name", and "1" under "flow".



USA - Mexico example

The table is sortable, by clicking on the table headings. For instance, the flows with the highest unit values can be seen by clicking on "unit\_value": the default behaviour is to sort in an ascending order, thus you should click again in order to have a descending order.



Sort

### Corrections

This is the "Corrections table", i.e., the table that contains all corrections that should be applied to the trade data.

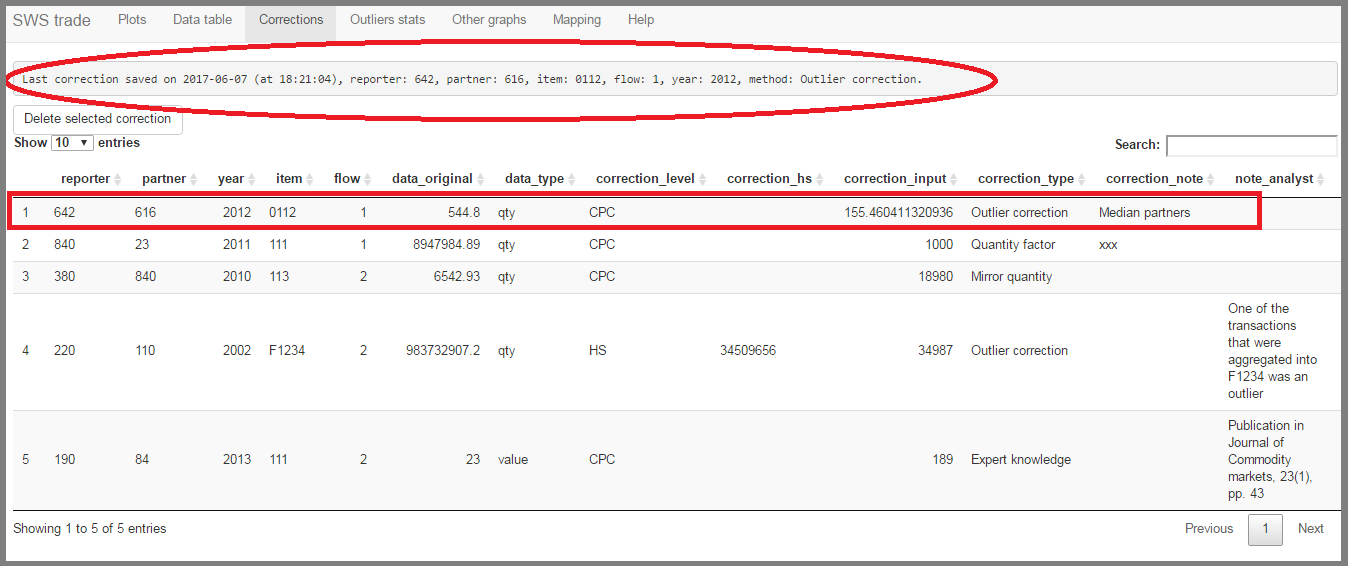
The columns are:

* **reporter**: Reporter country (M49 code).
* **partner**: Partner country (M49 code).
* **year**: Year.
* **item**: Item (CPC code).
* **flow**: Flow (1 = import; 2 = export).
* **data\_original**: The data originally produced by the module.
* **data\_type**: The type of data (qty = quantity; value = monetary value) -- **Currently only "qty" is supported**.
* **correction\_level**: The classification code to which the correction is applied -- **Currently only "CPC" is supported**.
* **correction\_hs**: If "correction\_level" is "HS" in this field should be indicated which HS code needs the correction.
* **correction\_input**: The corrections to be applied.
* **correction\_type**: See the "Choose a type of correction:" options in the "Plots" component.
* **correction\_note**: Automatically inserted string; Varies automatically by "correction\_type".
* **note\_analyst**: The note inserted in the "Optional comment/note:" field in the "Plots" component.
* **note\_supervisor**: Supervisor note -- **Not yet implemented**.
* **name\_analyst**: Name of the person who inserted the correction
* **name\_supervisor**: Name of the person who approved the correction -- **Not yet implemented**.
* **date\_correction**: Date when the correction was saved.
* **date\_validation**: Date when the correction was approved -- **Not yet implemented**.

You can delete a correction by clicking on it in the table (it will be highlighted) and then clicking on the "Delete selected correction" button above the table. An example will be shown further.

Once you save a correction you should see a message like the following:

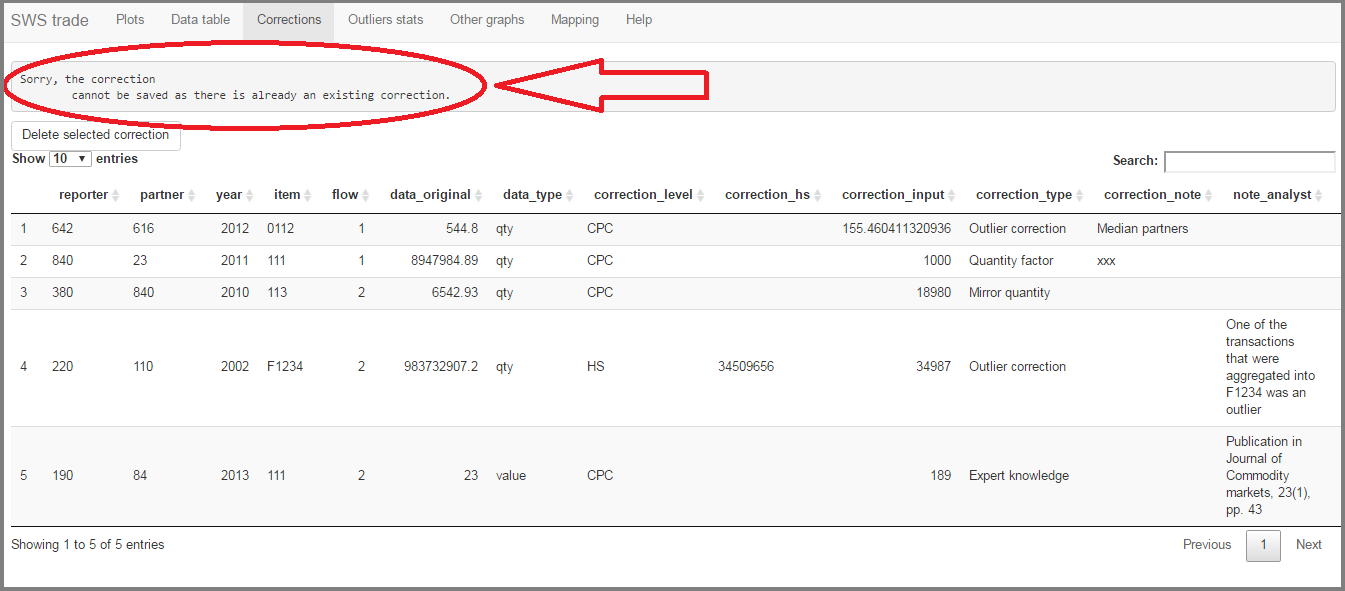
in the box above the corrections table if the saving was succesful. This is also shown in the next figure.



Correction confirmation

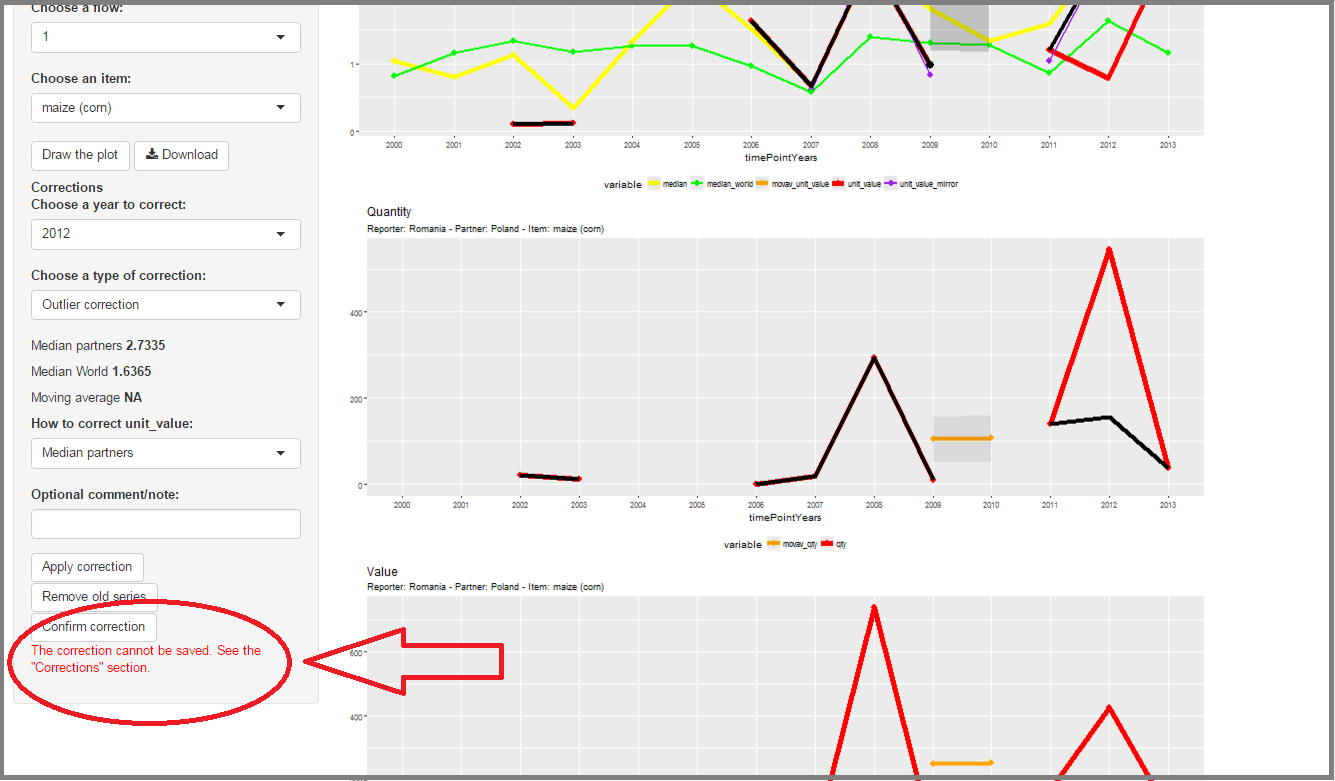
If the correction could not be saved because there is another correction for the same flow, the message will be:

This situation is shown in the next figure.



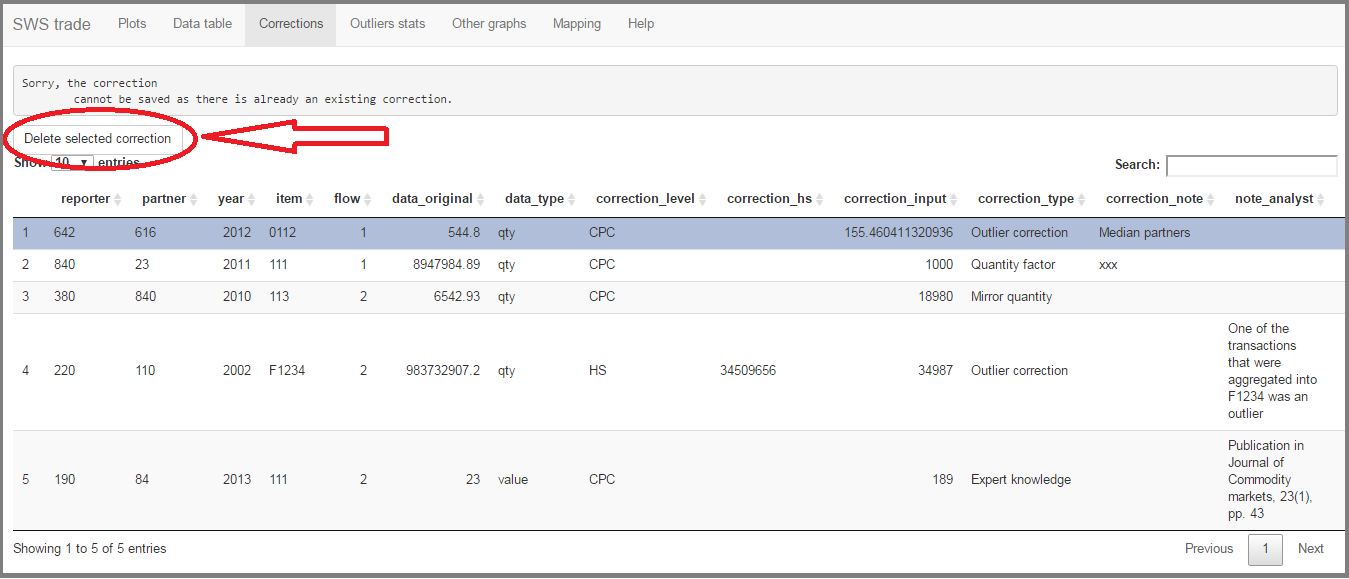
Correction failure message

You will also have an immediate feedback in the "Plots" component if the correction could not be saved. The message says "The correction cannot be saved. See the "Corrections" component." See next figure.



Correction failure message in "Plots"

In this case you should think about which correction you really need/want to retain. If you want to overwrite a previous correction in order to save a new one, you have to delete the previous one (as explained above).



Delete correction

**This table should be saved in the SWS. Currently, this is not yet implemented.**

### Outliers stats

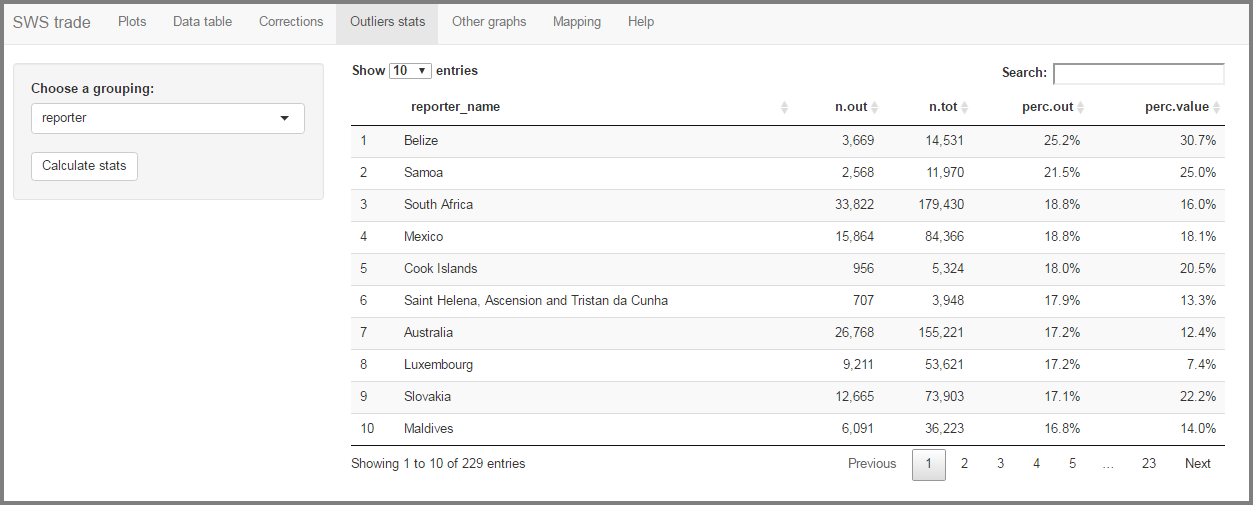
In this component some statistics on outliers will be shown (the criterion chosen in "Datatable" will be used; by default it is "Fixed threshold"). The aggregations that can be done are by:

* reporter
* item
* reporter and item

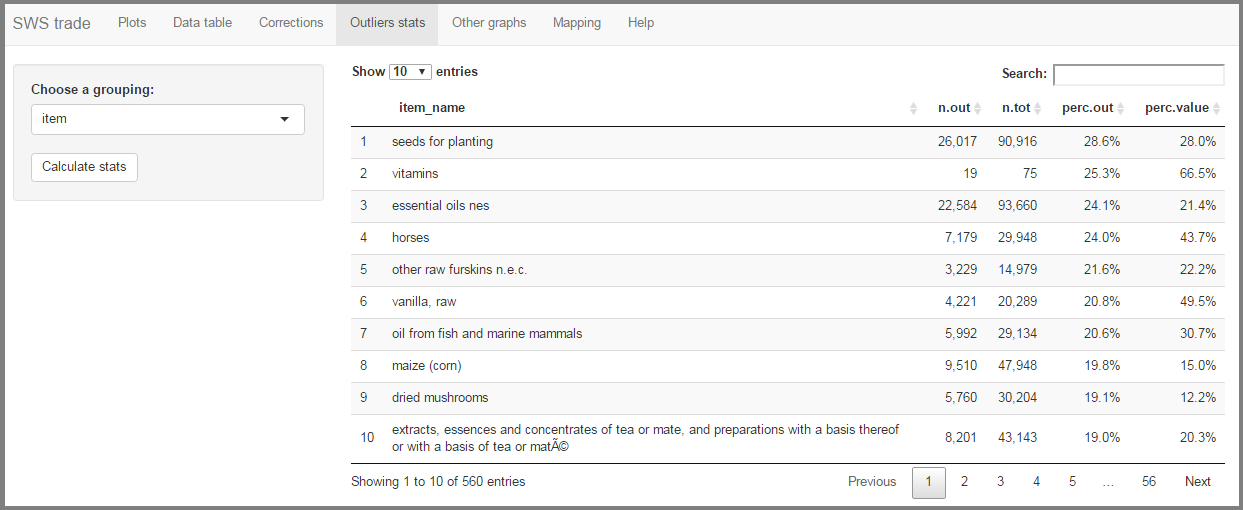
The table reports:

* **n.out**: Number of outliers.
* **n.tot**: Number of total observations.
* **perc.out**: Percentage of outliers (n.out/n.tot).
* **perc.value**: The percentage of the value of observations considered outliers over the total value.

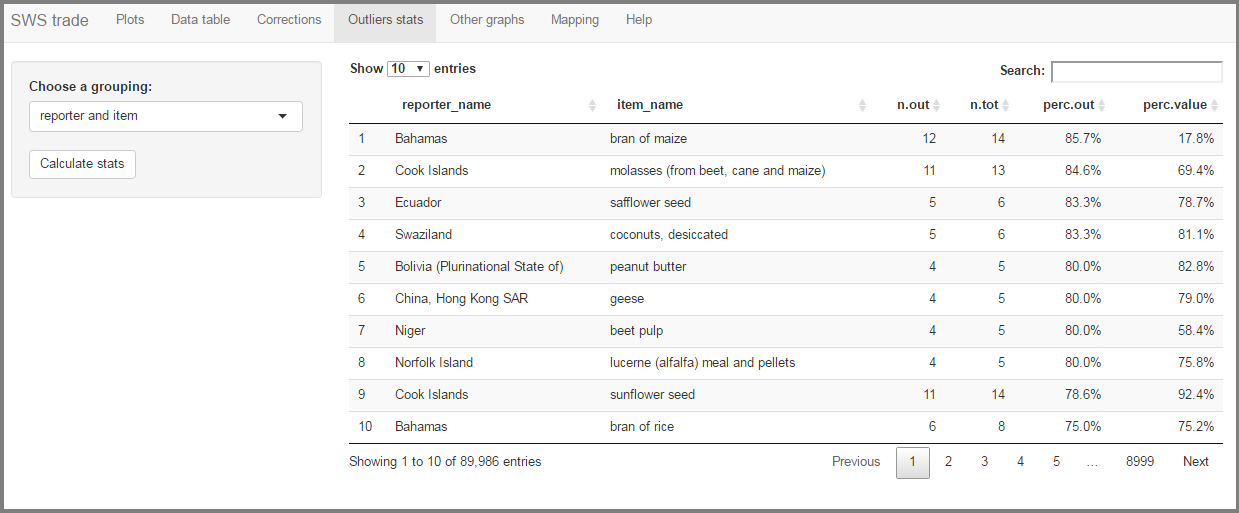
The following graphs show the tables that can be generated.



Outliers by reporter



Outliers by item



Outliers by reporter and item

### Other graphs

This component is intended to have some graphs that helps the analyst to spot the cases that need more attention (the outlier criterion chosen in "Datatable" will be used; by default it is "Fixed threshold"). Currently it has the following graphs (click "Generate graphs" to see them):

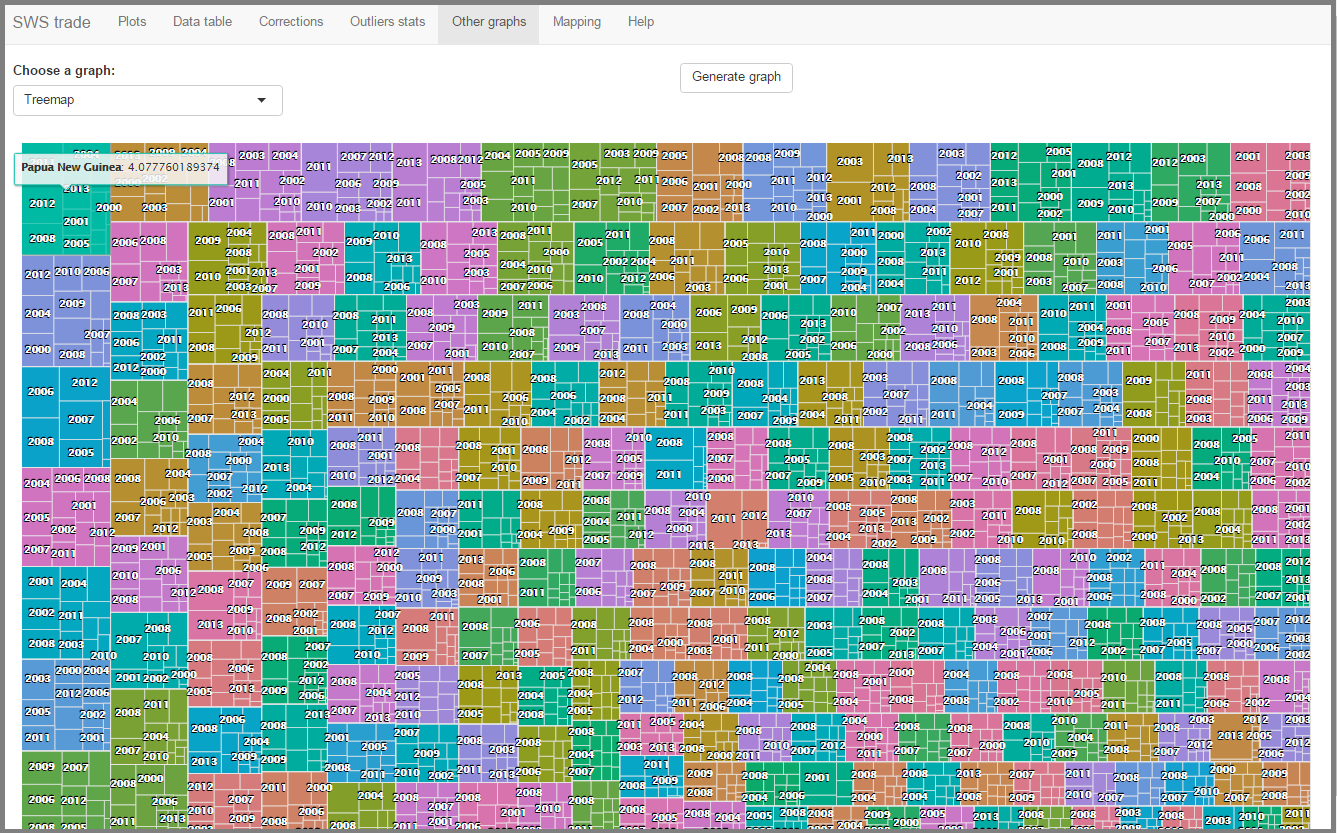
* **Heatmap**: By reporter and year, different colors show the distribution of outliers. See the legend under the graph.
* **Treemap**: This graphs show a hierarchic distribution of outliers: the first rectangle shows the country ranked first by percentage of outliers and within this rectangle it shows rectangles with years ranked by percentage of outliers; the second rectangle shows the country ranked second by percentage of outliers, and within this rectangle it shows rectangles with years ranked by percentage of outliers; repeat for all reporters.

The next figure shows how the heatmap looks like.



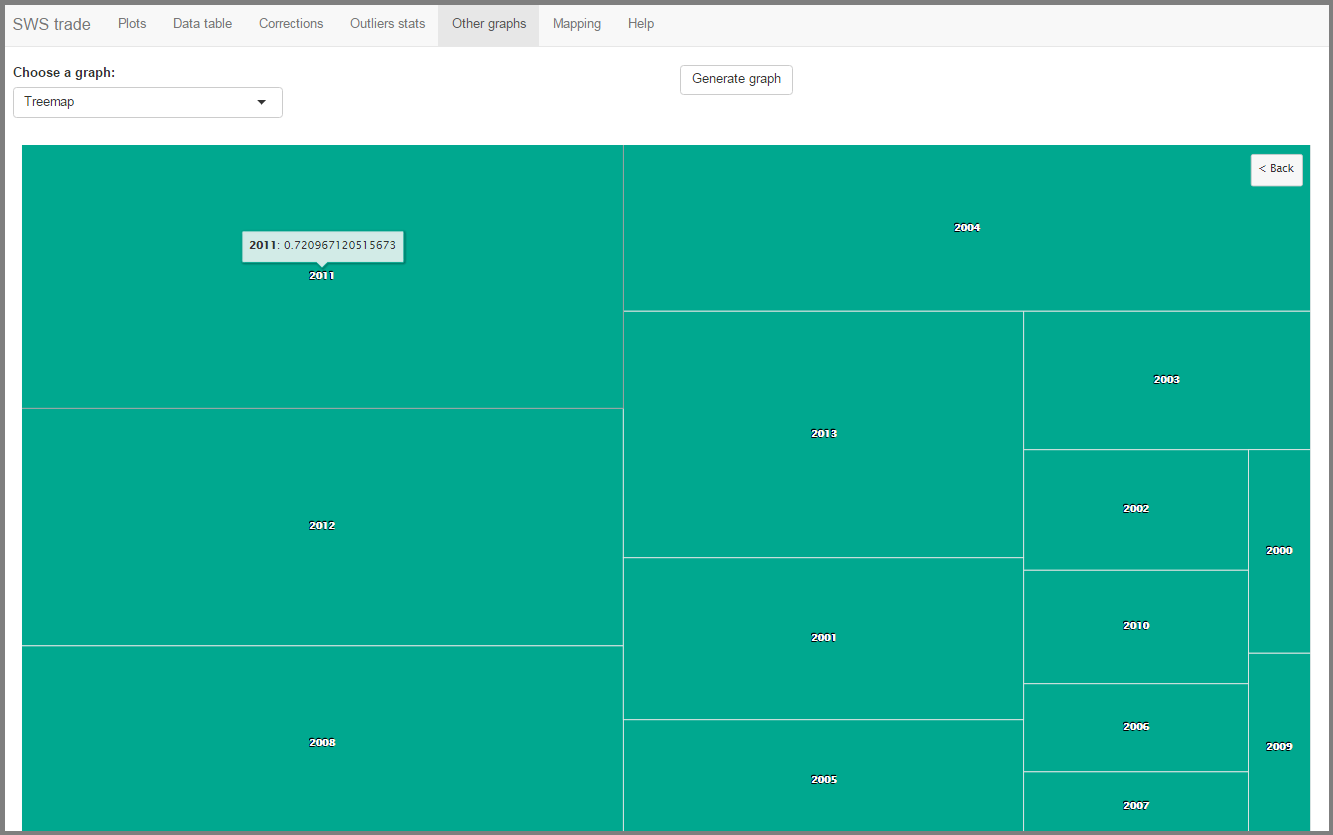
Heatmap

The next figure shows the initial screen of the treemap. As said before, this is a hierarchical visualisation: in the upper leftmost square the reporter with the highest percentage of outliers is displayed, below you find the second, then the third and so on.



Treemap

If you click on a square, it will show the distribution of outliers by year ordered from the most problematic year (you can go back by clicking the "< Back" button placed in the top-right of the graph). An example is shown in the following figure.



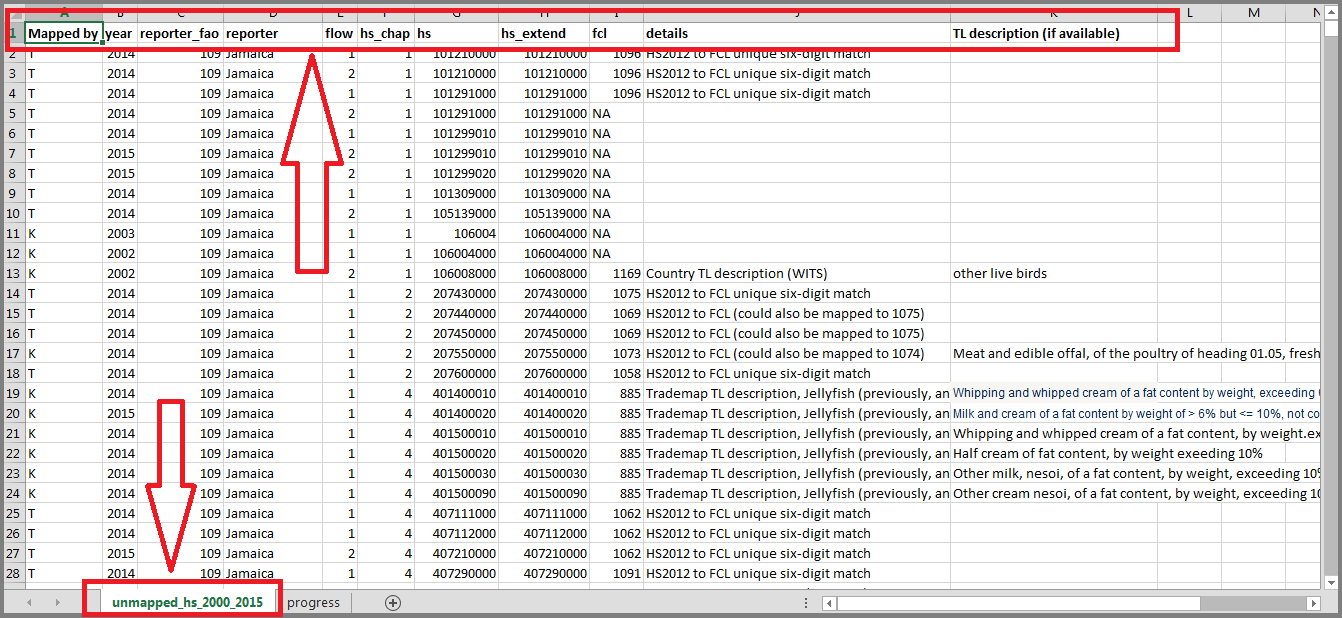
Treemap, country distribution

### Mapping

**NOTE: this component is not being used as the mapping has been integrated into the complete\_tf\_m49 module.**

This component helps you to transform an XLSX file with mapped HS>FCL codes into a CSV file of mapped codes to be uploaded into the SWS.

Your XLSX file should be structured like shown in the following figure.



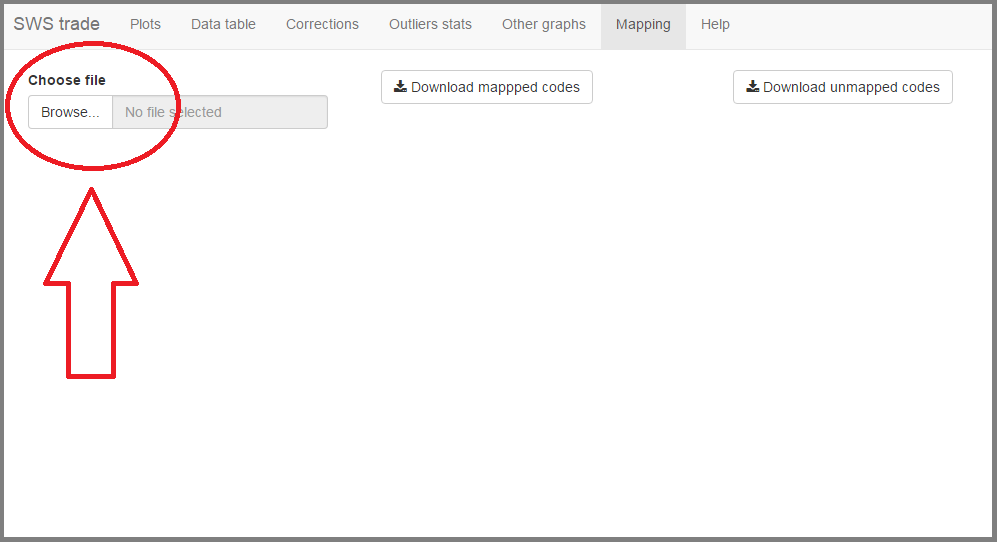
Mapping file

The following columns need to be present in the **unmapped\_hs\_2000\_2015** sheet:

* **Mapped by**: An optional string that indicates who did the mapping
* **year**: the year where the code was first found to be unmapped
* **reporter\_fao**: the FAO country code of the reporter relative to the unmapped code
* **reporter**: the name of the reporter
* **flow**: flow of the unmapped code (1 = imports; 2 = exports)
* **hs\_chap**: chapter of the **hs** code
* **hs\_extend**: extension of the **hs** code to the maximum HS length code used by the country for that year
* **fcl**: the FCL code corresponding to **hs**

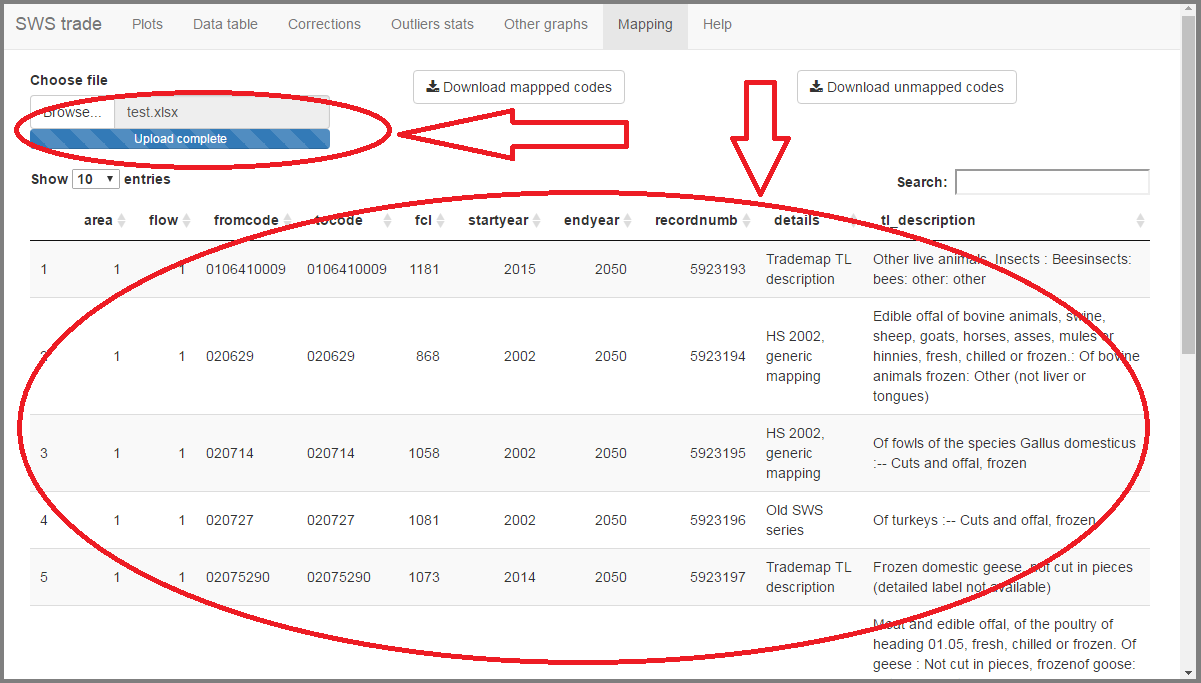
In the previous figure you will notice that some HS codes were already mapped to FCL and that some of them are not yet mapped (there is an **NA** in the corresponding cell).

You can process this file by clicking on "Browse..." as shown in the next figure.



Upload file

You will get a confirmation that the file was successfully uploaded when the "Uploaded completed" message appears below the button and a table of mapped codes appears, as shown in the next figure.

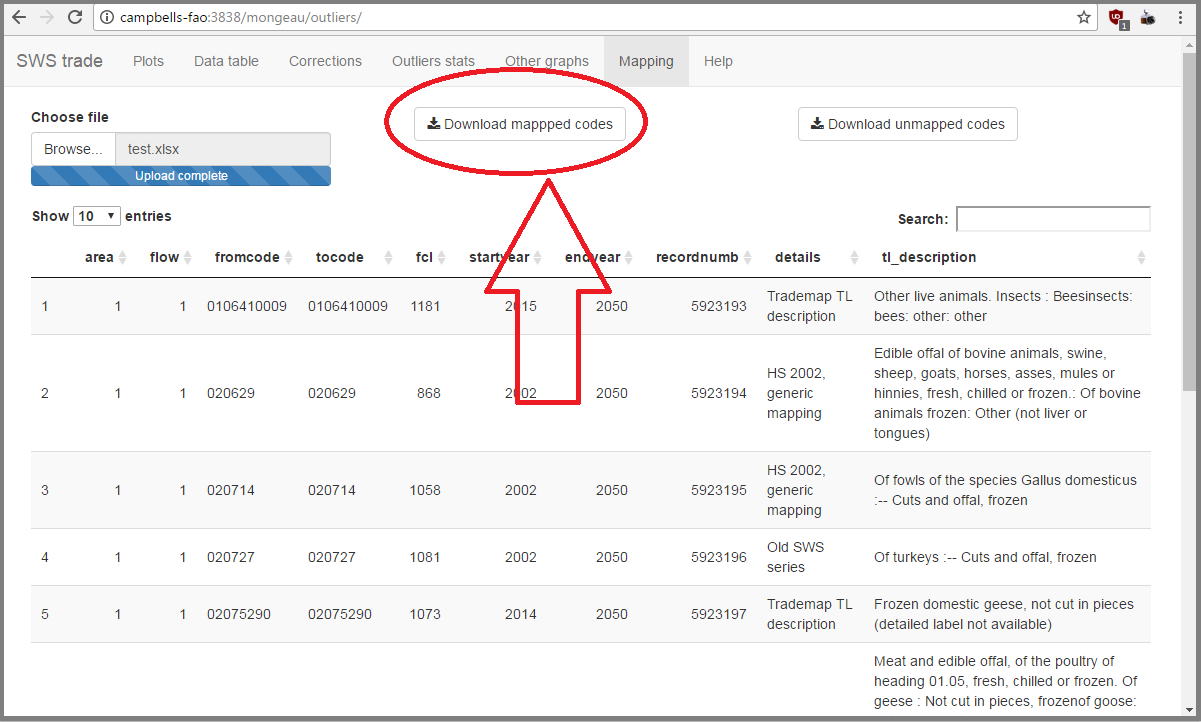


Upload successful

The tool will take care of doing three things:

1. it will map unmapped codes that can be mapped by using a generic 6-digit HS code (available in the following file: <https://github.com/SWS-Methodology/faoswsTrade/blob/master/data-raw/HS2012-6%20digits%20Standard.xls>
2. it will convert the original format of the file into the format used in the mapping table used by the trade module (e.g., it will generate a **startyear** and **endyear**) by combining the manually updated codes with the automatically updated codes (see previous point)
3. it will generate a file with the remaining unmapped codes

The CSV file of mapped codes ready to be uploaded into the SWS can be downloaded by clicking the "Download mapped codes" button, shown in the next figure.



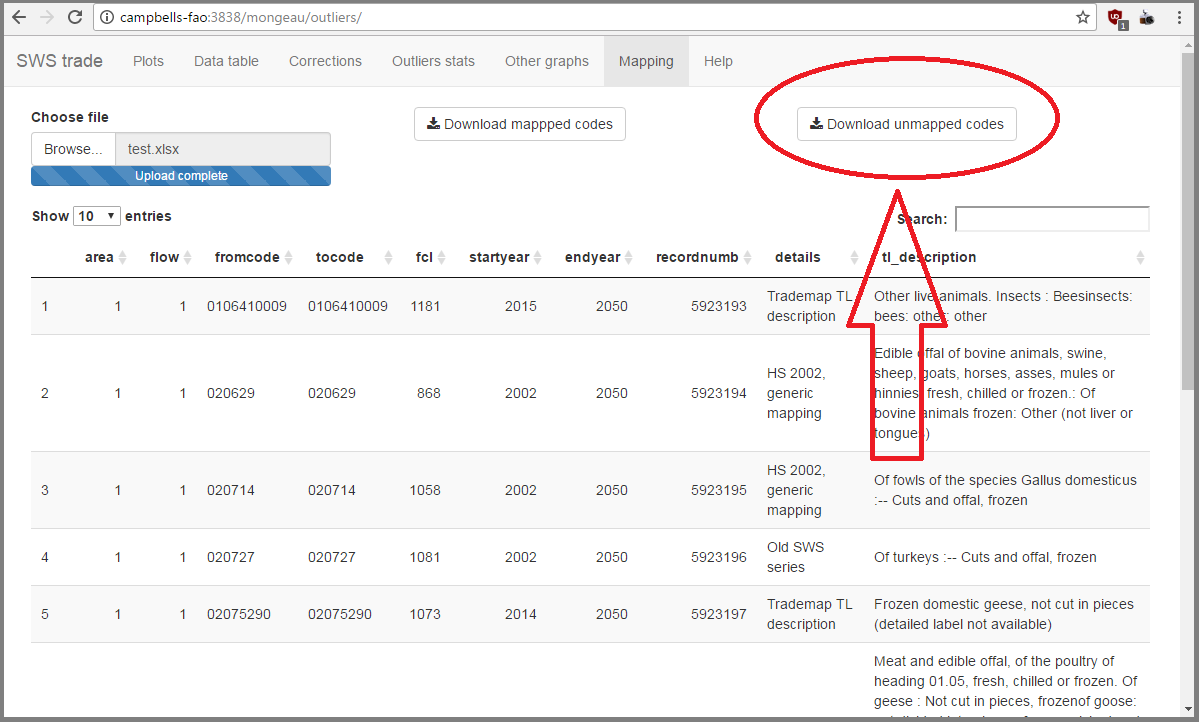
Download mapped codes

The downloaded file will have the following columns:

* **area**
* **flow**
* **fromcode**
* **tocode**
* **fcl**
* **startyear**
* **endyear**
* **recordnumb**
* **details**
* **tl\_description**

**Note**: this is not a file you are supposed to work on. If for some reason you open this file, do not save any modification. Store this file somewhere in your local computer, and send it to Team A in order to have it appended to the mapping file available on SWS.

The file with the remaining unmapped codes can be downloaded by clicking the "Download unmapped codes" button, as shown in the next figure.



Download unmapped codes

This file will be your new working file, i.e., the file that you will continue to work on in order to map the missing codes. Please open this file, save it as an XLSX file and continue with the mapping process in the .xlsx file. Once you are ready to have another batch of mapped codes translated into the file that will be appended to the mapping file on SWS, start the process again (upload the XLSX file, let the tool map automatically some codes, store the CSV of mapped codes somewhere on your PC and send it to Team A, and finally continue the loop with a new working XLSX file).

## Module/FAOSTAT comparison

The "flows" module can be accessed at <http://hqlprsws1.hq.un.fao.org:3838/flows/>.

### Rationale

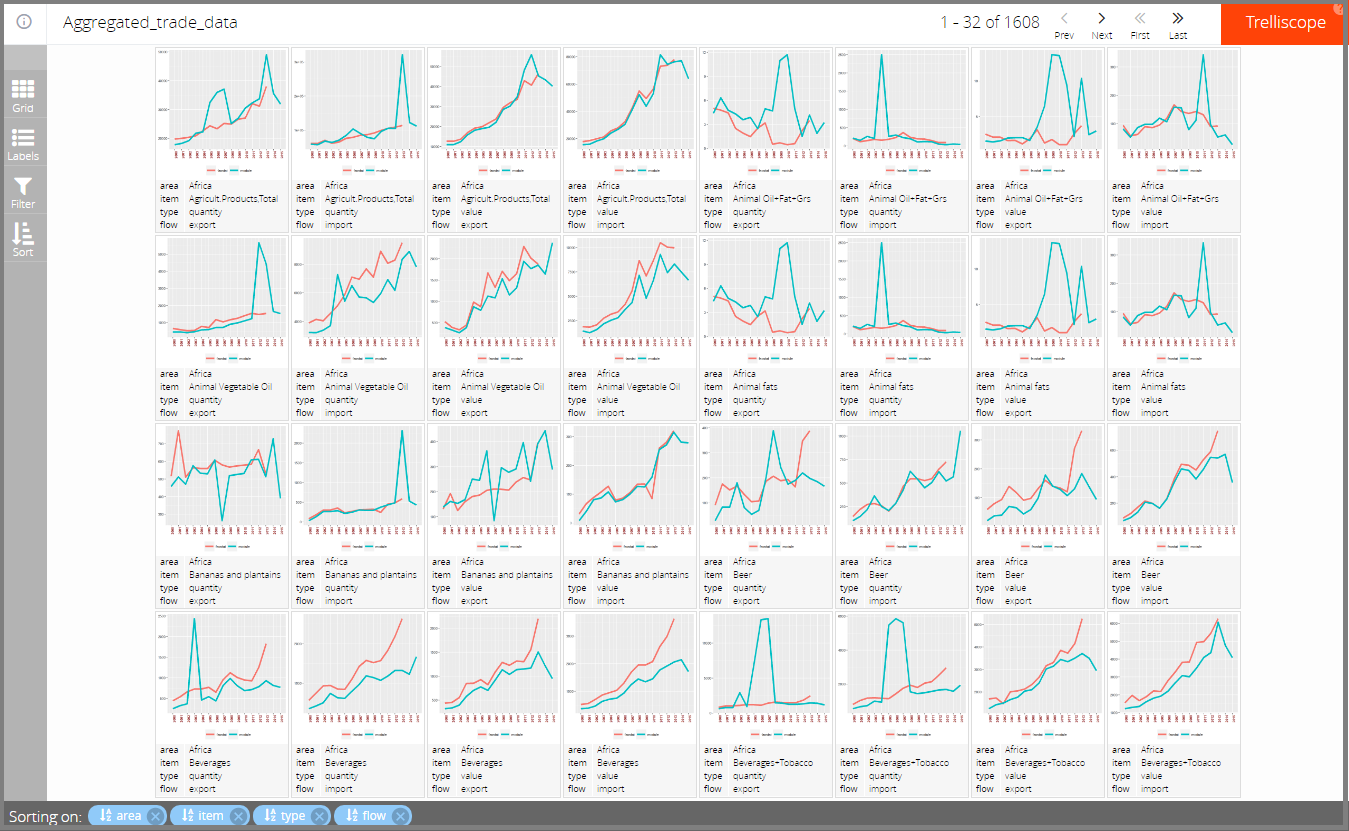
Before showing in detail the tool, an explanation about it existence is necessary: it allows to compare, with a handy visual presentation, the results from the new trade module with previously published FAOSTAT data. Given that FAOSTAT data underwent years of validation, it serves as a validation benchmark for the new module data. Discrepancies can be spotted very easily, allowing this way to validate new results.

A key point in the tool is that it presents "small multiples" of the time series of module and FAOSTAT data. What a "small multiples" means is explained at <http://www.juiceanalytics.com/writing/better-know-visualization-small-multiples>.

### Initial screen

When you open the tool, it presents one main section composed of various small graphs representing the time series of commodity/country aggregates of imports and exports. The module data is given by the green line, while the FAOSTAT data is the red line (FAOSTAT data ends in 2013).

The initial screen can be seen in the following figure.



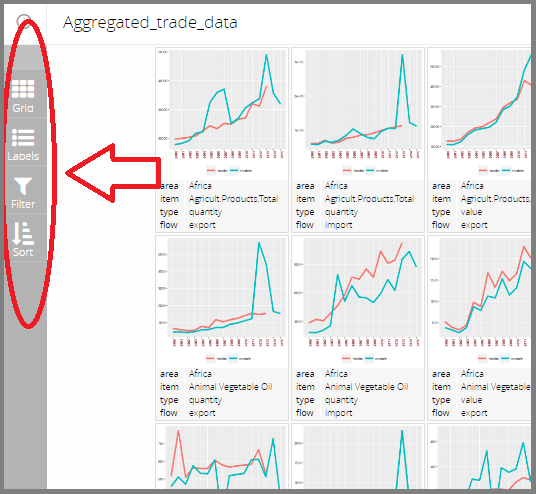
Initial screen

It is possible to scroll the graphs by pressing the buttons at the top of the page, as shown in the next figure.



Scroll graphs

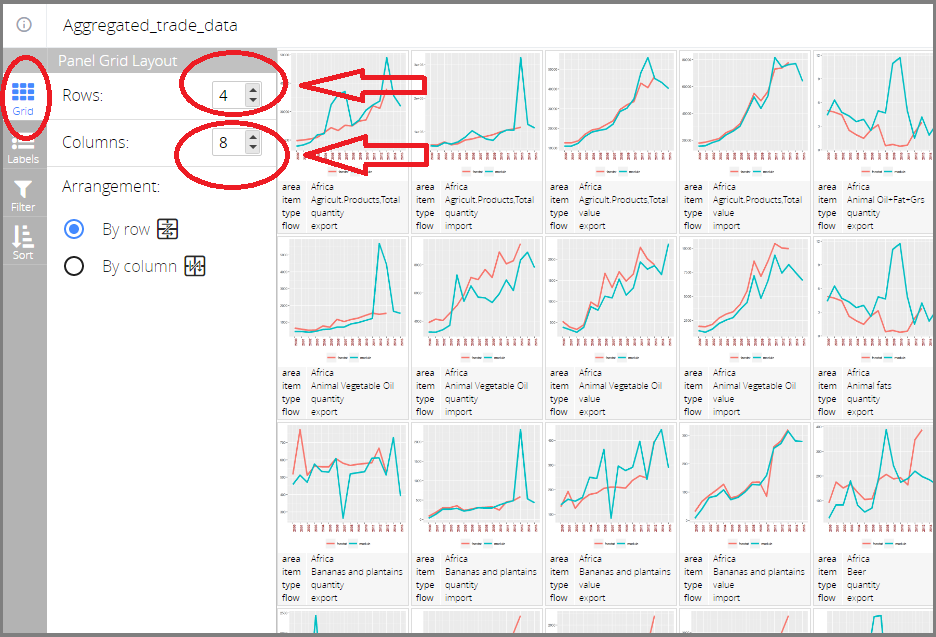
The left side menu is very important as it allows you to: select the number of graphs to display, select the filters to show, filter the graphs by specific conditions, sort the graphs.



Left side menu

#### Select the number of graphs to display

If you find that there are too many graphs in the screen, or, on the contrary, that there are too few, you can select in how many rows/columns you want to split the central screen to put the graphs. It can be accessed in the left side menu, as shown in the next figure, where the number of rows is set to four and the number of columns to eight (like in the figure of the initial screen).



Left side menu, Grid

#### Add/remove labels

If you know what the graphs represent, you may want to remove some labels so that the graphs increase a bit. For instance, in the first screen we know that all the graphs refer to "Africa" so that if we remove this label from below each graph, we do not loose any information. Actually, in this case you should be careful as if you start scrolling the graph other country aggregates may show up and you will lose this information. A more useful application of this is when you use the "Filter" so that you know that you will condition the graphs to some specific condition and you will not take the risk to remove information. The "Filter" menu will be shown below. For now, let us remove the "area" aggregate from the initial screen. It can be donw by deselecting the desired label from the "Labels" menu, as shown in the next figure. Notice that under each graph the "area" label disappeared.



Left side menu, Labels

Of course, if you want to add a label, you have to select the desired and unselected label.

#### Filter by variables

An useful feature is the possibility to condition the graphs to some specific values of the variables. For instance, you may be interest in comparing module and FAOSTAT quantity data relative to imports in Europe. This can be achieved by selecting the "Filter" menu, as shown in the next figure.

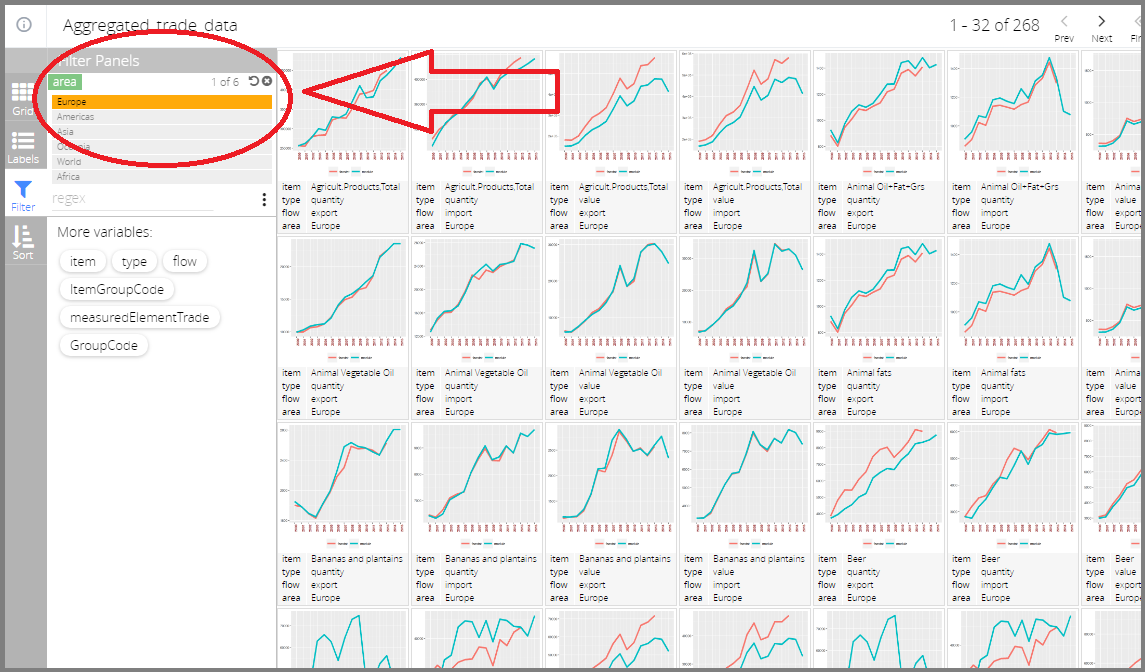


Left side menu, Filter

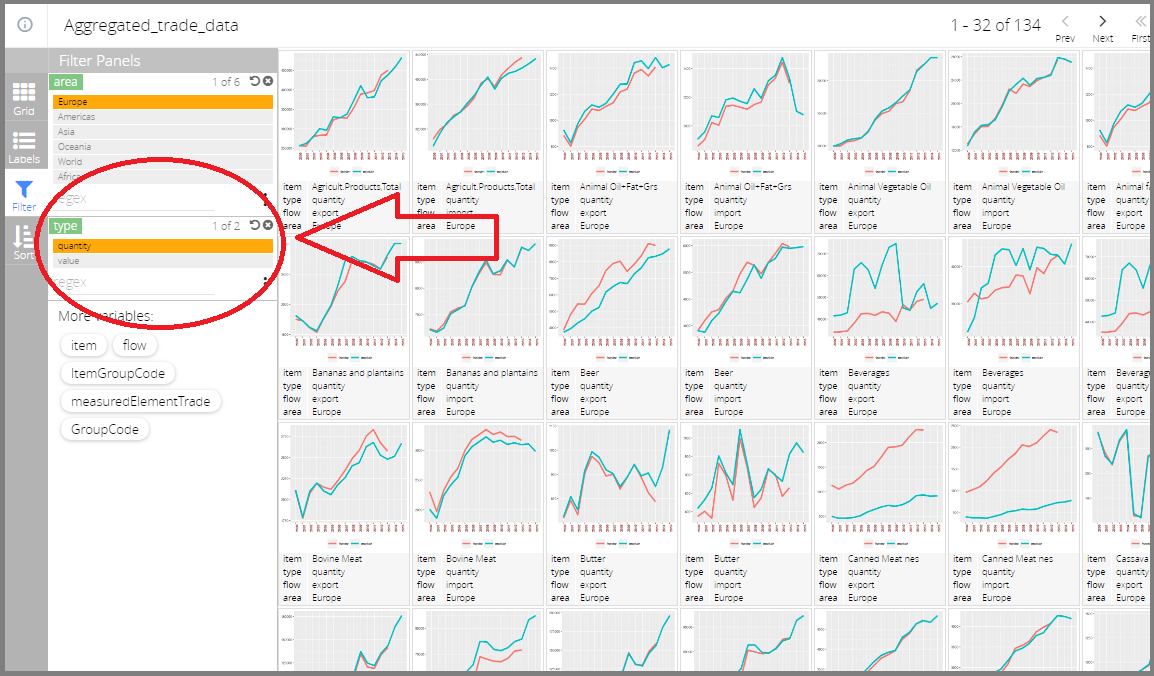
Once the menu is open, it shows the variables on which the conditioning can be done:

* area
* item
* type
* flow
* ItemGroupCode (item's numeric code)
* measuredTradeElement
* GroupCode (area's numeric code)

Following the example above, we can click on "area" and click on "Europe", then on "type" and click on "quantity", then on "flow" and click on "import". Note that the actual order in which you do the selection is not important. The filtering given in this example is shown in the next figures.



Left side menu, Filter by area



Left side menu, Filter by type



Left side menu, Filter by flow

The filtering by variables can be multiple. For instance, if you want to see Europe's and Asia's import quantities, you can select "Asia" and "Europe" in the "area" variable, as shown in the following figure.



Left side menu, Filter by multiple areas

#### Sort by variables

If you want to have plots in an order that follows some variable, you can do it by clicking the "Sort" menu and selecting the variable on which you want to sort the graphs.

Consider the results as they appear in the previous figure: Asia comes before Europe (increasing alphabetical order). If you want to have Europe before Asia, you can select the "area" variable in the "Sort" menu.