**Trade completion, validation and consolidation through the R-UNSD-API**

**June 2013**

**Adam Prakash**

1. **Introduction**

The reliability, consistency and accuracy of trade data affects the credibility of derived statistics and metrics upon which they are based. These include indicators that measure the volume or value of trade by a geographical entity, such as the world or sub-Saharan Africa or a single country; or that by an economic group, such as Least Developed Countries (LDCs). Another example for which the quality of trade data is a prerequisite is given by commodity-based accounts, such as Food Balance Sheets (FBS). In this example, the correct reporting of trade flows in expanding international markets is critical for measuring accurately food availabilities within a country.

Furthermore, the direction and composition of trade, termed the *structure of trade,* has important bearing on models that simulate policy shocks, e.g. AGLINK-COSIMO, CGE models, as trade structure determines how the effects of a given policy shock are distributed across economies throughout the world. Furthermore, that FAOSTAT can be an important source of data for these classes of models, ensuring that trade data meet the highest standards of quality would also greatly enhance the demand for FAOSTAT.

A corollary of the above is that discrepant trade data would lead to the erroneous compilation and derivation of such statistics, and what is more, rendering inference to policy-making unreliable.

In order for trade data to be meaningful and informative, fulfilling the “evidence base”, there are formidable methodological issues that require addressing. These include non-reported outflows or inflows, suspect unit values which cast doubt on the quality of either reported quantities and values, and cases in which both the importer and exporter have notified a non-zero trade flow but the notifications do not match within acceptable bounds.

A prior attempt at reconciling unbalanced trade in the FAO Statistics Division using a systematic approach did not succeed, as the researchers did not fully grasp the intricacies of trade. For instance, all zero trade notifications were treated as missing data, regardless if the country did not engage in the international market transactions. It was also presumed that trade data are strongly determined by their past, which led to the smoothening of discrete jumps. On both counts, there are strong reasons why countries do not engage always in trade and also why trade could be volatile year-to-year. The upshot was that the “balanced” trade file bore little resemblance to reported trade flows, and in many instances empirically reliable data were dismissed and subjected to imputation.

The challenge then is to apply a set of orderly rules that condition on all of the available information contained in trade data, which do not introduce systematic bias in the results. In doing so, enormous efficiency gains are to be had in terms of workflows and similar improvements to the quality of trade data disseminated. Drawing on established peer-reviewed methods, including those applied to balance trade in the GTAP database, this paper proposes a set of efficient rules for the completing, validating and consolidating trade.

Finally, the grounds and justification for a new approach are not discussed in the present paper but provided in the following papers:

* towards a new approach for the collection, processing and dissemination of trade data (see DCWM TRADE ImputationValidation\_v1 1\_Rev\_02Jul12.doc );
* classification (HS to FCL) discrepancies (see FCL-HS\_Item\_Analysis.doc);
* the extent of imbalanced and discrepant trade in FCL (see DQA FAOSTAT Trade ver0.doc);

1. **Data quality issues**

*A priori* there are many factors that affect the quality of trade data, and are currently documented by way of disclaimer in FAOSTAT.

1. time lag: an export reported in December of a given year could reach its destination in the following year (and would only then be reported as an import) giving rise to mismatched annual flows
2. misclassification of the product between the exporter and importer
3. exported quantities could be lost en route owing to accidents, weather conditions, etc. or redirected
4. recording errors by the reporting country
5. data confidentiality
6. place of origin/final destination inconsistencies
7. customs tax avoidance by misrepresenting a commodity on import or not reporting a transhipment. In particular cases: exports are not declared to circumvent an embargo or tax

The list can be further examined to identify sources of error – (ii), (iv) and (vi) certainly fall into this category. As for the others, information can used to infer missing flows, e.g. (iii), (v) and (vii). As for (i), a 5 percent rule degree of tolerance between the matching of flows can be employed.

However, another source of error that is not brought to the attention of users of FAOSTAT trade data pertains to the attempt at reclassifying products from the widely used HS (Harmonized Commodity Description and Coding System) to FASOTAT Commodity List (FCL). The potential for errors to be introduced is evidenced by the fact that only one-third of HS products have a unique mapping to FCL. This issue is summarized in “HS-FCL\_1-to-1.xls”, as well as in a forthcoming paper.

1. **Collecting trade data**

Trade data are gathered seamlessly through UNSD web services, either though SDMX or element-based-XML. The latter service was invoked given that FAO-ESS has no IT capacity to process SDMX structured data. The following browser command was generalized in java to automatically download and process UNSD data:

[http://comtrade.un.org/ws/getTariffLine.aspx?cc=*j*&comp=false](http://comtrade.un.org/ws/getTariffLine.aspx?cc=j&comp=false)

Parameters and access points for UNSD web services are provided in Annex A.

1. **Processing trade data**

**Assumptions**

* The quality of value data is presumed to be superior than for traded quantities. This is evidenced by the greater availability of such data *vis a vis* quantities. As tariffs and other import duties are typically charged on the value of merchandise received at the border, there would conceivably be better monitoring systems by officials.
* For non-reporting exporters whose flows are not notified by any importer, there would be no information available within the dataset to estimate such flows. Such scenarios are presumed extremely improbable.
* It would be a broad over-generalization to assume that trade flows are conditioned by their own past. For some commodities and for some countries, such as import-dependent and export-driven economies this may be the case, where regular shipments are received/delivered from/to one destination. As a consequence, only contemporaneous information is employed to complete missing values.
* Price volatility is a permanent feature of international markets. It would therefore be misleading to disregard trade flows possessing unit values that undergo substantial annual variability.
* In a similar vein, substantial heterogeneity exists within product groups at the most disaggregated level of standard reporting in the HS system, namely at the six digit level. This heterogeneity is reflected in wide ranging unit values reflecting variations in product quality, grades, characteristics and levels of processing. The degree of heterogeneity needs to be acknowledged before dismissing flows based on seemingly errant unit values

**UNSD data structure**

Annual trade data for each commodity, *j*, obtained via the UNSD API, are structured by trade “regimes”: imports, *m*, exports, *x*, for each transaction *i* by reporting country *a* to partner country *b,* recording the quantity of trade, *q* and the value of trade, *v*. Or, more succinctly  and .

The data were then restructured, or in R terminology “reshaped” to produce a set of unique records. In addition, EU intra trade flows were discounted, conditional on the membership and accession of member countries for the reported year.

Using notation applied in the domain of propositional logic, *hypothesis**conclusion*, a set of data transformation rules were applied in the following sequence.

* **Data cleaning: self-trade**

***Rule1***:

for . That is, in instances when a country reported itself as a partner to exports or imports, records were subsequently eliminated.

* **Completion through mirroring**

The following rule was used to complete missing data when counterpart reporter/partner data existed

***Rule 2***: 

. Essentially, this rule eliminates the sparseness in the trade data set, by completing non-reported flows with counterpart reported flows. However, as mentioned previously, for non-reporting exporters whose flows are not notified by any importer it would not be possible to estimate such flows from within the dataset. The likelihood of such a scenario is presumed extremely improbable.

* **Completion through unit values: missing quantities but values exist**

Import and export unit values were calculated as follows:  along with median/modal[[1]](#footnote-1) unit values for all imports  and all exports. While the median may not actually occur in the distribution, it is the "balance point" of the distribution, being unaffected by outliers as is the popular measure of central tendency, namely the mean. Likewise, the mode is little affected by outliers. Indeed, distributional plots of the unit values pointed to significant skewness and the choice of using measures other than the mean is appropriate.

***Rule 3***: 

Accordingly, missing quantities are derived by the quotient of trade value over the respective unit value.

* **Validation through unit values**

The following rules are applied to identify rogue unit values that necessitate adjustments to quantities.

**“*Order of magnitude*”**. Misplaced decimal points can give rise to such errors, and are corrected by testing whether unit values are of a higher order of magnitude than the average unit value.

***Rule 4:*** 

for *n*≠0.

“***Out of range***”: If unit values are either below or above the range of values for which non discrepant flows are observed, quantities are recalculated based on the modal/median unit value. From the set of near matching trade flows  , define  and as the maximum and minimum unit values, respectively, of the set of permissible values. The rule is then, if , that is the value share of near matching trade flows is at least 50 percent of total trade value of the commodity, then

***Rule 5 :*** 

“***Relativity***” : This rule relates to the departure of import unit values from export unit values. A ratio of more than three is presumed to capture this presumed anomaly.

***Rule 6:*** 

There appears no plausible reason why import unit values (*cif* basis) would be lower than export unit values (*fob* basis), but cursory inspection reveals occurrences that are mostly confined to minor transactions.

* **Reliability**

There are no rules that govern the coordination of reporting trade, with outflows independently reported of inflows. This is perhaps the major reason for the large percentage of discrepant transactions found in officially reported trade data. Using Gehlhar’s (1996)[[2]](#footnote-2) analogy of ballroom dancing, the problem to be confronted and eventually resolved can be illustrated as follows:

Consider what would happen if there was no communication between dance partners. Each dancer attempts to lead regardless of whether their partner knows the dance. It would only be when *both* partners know the dance very well when the couple is *in-step*. Even if one partner knows the correct dance steps the couple painfully stumbles over one another *out-of-step*. And so it is for trade reporting. What can be done with a non-communicative group of ballroom dancers? Rules must be imposed. One rule is to have all women follow exactly what the men do, analogous to choosing reported imports over reported exports. This would at least help to keep partners *in-step* with each other. But it gives no assurance that the correct dance is being performed. Not all men know the dance steps for every dance. And the same goes for women. Another solution is to work out a compromise. Women could lead the first half of a given dance and men lead the second half. This is analogous to taking an average of the two reported values of trade. But there is no reason to believe that the average is the actual trade flow. But the objective in ballroom dancing is not simply to have everyone *in-step* with each other, but to perform the correct dance *in-step*. The same goes for bilateral trade where the correct trade flow is most important. It seems only logical that the dancer who knows the dance should lead regardless of gender. This approach makes use of all dance talent found in the entire group, and is the approach taken in reconciling trade data.

Misreporting of bilateral trade can either be intentional or unintentional, but the reasons for aberrant reporting are not important. A common problem encountered in bilateral trade concerns partners who are not correctly identified or commodities that are not correctly classified. The erroneous identification of partners leads to the under-reporting of flows for one partner and over-reporting for another without changing overall trade. It is for this reason that using totals reported by countries provides little or no insight in determining reliability, rather through assessing the integrity of trade flows.

Yet, there are a number of ways that reliability could be measured. One way is measure the average discrepancy across bilateral trade flows for each reporter. But size of discrepancy can diverge significantly and is not necessarily a predictor of reliability. As Gelleher (1996) states, “it is like judging a dance partner based on just how bad their worst partner dances were. Conventional wisdom would suggest that it is the rate of successful dances that reflects knowledge of dance talent…Accordingly, the sum of the total quantity of *accurate* partner trade as a share of total reported trade is well suited for measuring reliability.” With differences, potentially wide, between the *cif* and *fob* basis, quantities are used instead of values.

In the first instance, what is required is to identify those countries which systematically misrepresent their reported bilateral trade. In this regard, an overall error metric is calculated for each transaction over those that have not been modified by prior rules, . The direction of error is not important, but the magnitude is, and also in the absolute sense:



Next, reliability indices are calculated for importers and exporters based on the quantity of reliable transactions, for which discrepancy as mentioned earlier is set at an arbitrary 5 percent over total transactions:



where  denotes those transactions for which .

The reliability index for all reporters and partners is then adjusted to eliminate the transaction which produces the largest error. Through removing the largest discrepant transaction, the reliability of better countries is improved much more relative to poor reporting countries.



The reliability rule is now straightforward:

***Rule 7***: 

, in that for any mismatched trade flow, the more reliable quantity reporter replaces the counterpart transaction.

1. **Metadata and logging**

For every trade flow/transaction that has been subject to transformation, the nature of the process is flagged. Moreover, if the quantity of trade that has been subject to modification exceeds 1 percent of global trade for the commodity in question, the details of the transaction are logged, prompting further research and possibly manual intervention after verification.

1. **Selected preliminary results**

The rules were implemented in R with a java-based routine to invoke the UNSD-API. The time taken to harvest the data, and to process via the rules, amounted to around 20 seconds per commodity.. Assuming approximately 700 HS agricultural-related products, for which trade data are required to be collected and processed, the total time theoretically stands at around 4 hours. This compares to the 2000 labour hours[[3]](#footnote-3) currently involved in preparing trade data.

Three diverse commodities were chosen to test the new rules-based approach:

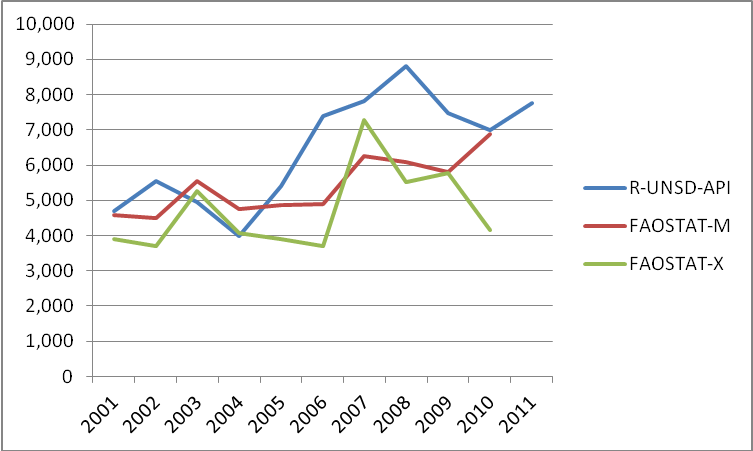
* Wheat (grain) – HS 1001, FCL 0015
* Vanilla – HS 0905, FCL 0692
* Grapes, fresh – HS 0806.10, FCL 0560

The results are shown below together with comparable data from FAOSTAT and the USDA.

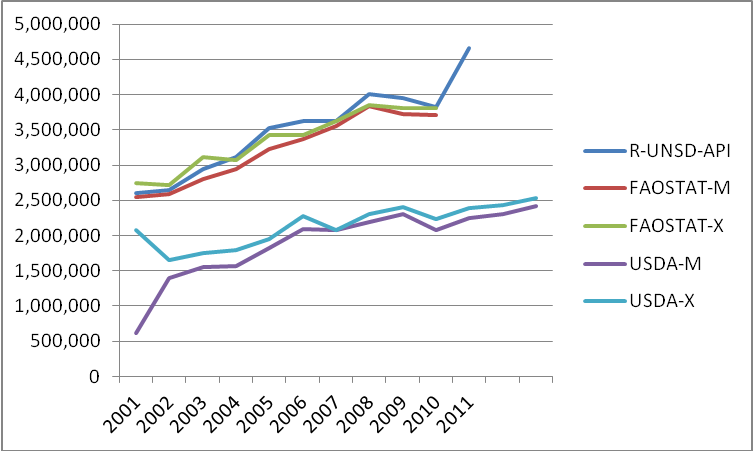
**Global wheat grain trade (M=imports, X=exports; 1000 tonnes)**



**Global vanilla grain trade (M=imports, X=exports; tonnes)**



**Global grape trade (M=imports, X=exports; tonnes)**

****

Two cases in point:

* **Australia’s wheat grain exports**: for confidentiality reasons, the sizeable wheat exporter does not disclose quantity information nor on where its wheat was destined. However, this information was obtained from partner notifications and also verification of total trade value with unit values.
* **Madagascar’s vanilla exports:** Madagascar reported 6 million kg of vanilla exports in 2003 to China. However, China did not report any such exports from Madagascar. Based on the unbiased export unit value, an order-of-magnitude error (a factor of 10) was identified in Madagascar’s notfification. Madagascar’s exports were revised down to 600,000 kg – incidentally, in line with past years.

1. **Conclusions**

This paper has provided a description of the rules-based appraoch to the trade work. The results are promising, not only in bringing consistency, plausability and consolidated flows in trade data but also in terms of its efficiency in gathering and processing data.

**Annex A: Parameters and access points for UNSD web services**

|  |  |
| --- | --- |
| **Regime Codes** | |
| 1 | Imports |
| 2 | Exports |
| 3 | Re-Exports |
| 4 | Re-Imports |

|  |  |  |
| --- | --- | --- |
| **UN Comtrade Code** | **WCO Abbreviation** | **Description** |
| 1 | - | No quantity |
| 2 | m2 | Area in square meters |
| 3 | 1000 kWh | Electrical energy in thousands of kilowatt-hours |
| 4 | m | Length in meters |
| 5 | u | Number of items |
| 6 | 2u | Number of pairs |
| 7 | l | Volume in liters |
| 8 | kg | Weight in kilograms |
| 9 | 1000u | Thousand of items |
| 10 | U (jeu/pack) | Number of packages |
| 11 | 12u | Dozen of items |
| 12 | m3 | Volume in cubic meters |
| 13 | carat | Weight in carats |

|  |  |
| --- | --- |
| **Commodities and Countries** |  |
| CmdCode: | Commodity code |
| CmdDesc: | Commodity description |
| IsLeaf: | Basic code without children |
| Parentcode: | High level of that commodity code |
| pfDesc: | Commodity classification |
| PfCode: | Commodity classification code |
| yr: | Year |
| rtCode | Reporter Code |
| ptCode: | Partner Code |
| qtCode: | Quantity code |

|  |  |
| --- | --- |
| **Access Points** | **Description** |
| *refs/getCommodityList.aspx* | Get list of commodity |
| *refs/getCountryList.aspx* | Get list of country |
| *refs/getDataAvailability.aspx* | Get data availability |
| *refs/getDATariffLine.aspx* | Get tariff line data availability |
| *refs/getExplanatoryNotes.aspx* | Get explanatory notes |
| *refs/getPublicationNotes.aspx* | Get publication notes |
| *get.aspx* | Get data in element-based XML format |
| *get.aspx?detail=true* | Get data in element-based XML format |
| *getSdmxV1.aspx* | Get data in SDMX version 1.0 format |
| *gettTotal.aspx* | Get Total Trade in XML format |
| *getConvFactor.aspx* | Get Conversion Factor in XML format |
| *getSdmxTariffLineV1.aspx* | Get Tariff line data in SDMX version 1.0 format |
| *getTariffLine.aspx* | Get Tariff line data in XML format |
| *getDataMapServer.aspx* | Get data in GML format |

1. The mode was calculated using the kernel density estimator:, with a Gaussian kernel, K, . In kernel density estimation, the single bandwidth *h* is replaced by *n* different bandwidths dependent on  A bootstrap of 100 resampled repetitions of the mode was also conducted to ensure precision, especially on smaller samples for which valid unit values exist, [↑](#footnote-ref-1)
2. Gelhar, M. “Reconciling Bilateral Trade Data for Use in GTAP”, 1996, GTAP Technical Papers, Purdue Technical Paper Series. See <http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1011&context=gtaptp> [↑](#footnote-ref-2)
3. Based on Team C’s 2013 work plan “Process the reported trade data for the year 2011, standardize and validate the trade data files (apron. 160 country files for 2011 and previous years)” in which 5 staff are envisaged to be engaged over 3 months. [↑](#footnote-ref-3)