UNITED NATIONS
DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS

STATISTICS DIVISION
ECONOMIC STATISTICS BRANCH

2019

Trade data processing and validation methodology

Joint paper of FAO[1], UNSD[2] and ITC[3]

Part: UN Comtrade

Main sections

- Detection and correction of outliers
- Estimation of quantities (reported outlier quantities, missing quantities)
- Aggregation methods (across commodity hierarchy)
- Validation checks (in conjunction with outliers' detection)

General information not uniquely fitting in one of the pre-defined chapters

UN Comtrade is the most used database for annual and monthly data on international merchandise trade statistics detailed by commodity and partner. UN Comtrade provides not only values for exports and imports, but also data on the (supplementary) quantity and (net/gross) weight, detailed by commodity and partner. For providing the most comparable and comprehensive information as possible, the quantity and weight information reported by countries is standardized and, where necessary and possible, estimated. The standardization for supplementary quantity units is based on the recommended units by the World Customs Organization (WCO) (see Table 1). However, UNSD allows a broader selection of units for the reporting countries (see Table 2). The UNSD Standard Quantity Units are introduced for capturing industry practices (e.g., hive for bees instead of items) and taking widely used national quantity units and international standard units, e.g., the SDMX-IMTS Code list[4], into account. The UNSD list is intended to be flexible, i.e., an extension is possible if needed.

In UN Comtrade, weight and quantity is reported as net/gross weight in kg, a quantity unit following the WCO and UNSD extended quantity units. Illustration 1 shows an extract from UN Comtrade for

[1] Food and Agriculture Organization of the United States
[2] United Nations Statistics Division
[3] International Trade Centre
[4] https://comtrade.un.org/sdmx

the commodity code 220110 which is water. Thus, the weight in kg is equal to the quantity measured in litres.

Illustration 1: Example for the dissemination of trade data in UN Comtrade[5]

Trade Flow	Commodity Code	Trade Value (US$)	Netweight (kg)	Qty Unit	Qty	Flag
Import	220110	$513,120,251	700,023,045	Volume in litres	700,023,045	0
Export	220110	$49,487,840	132,648,038	Volume in litres	132,648,038	0
Re-Export	220110	$5,772,863	5,694,317	Volume in litres	5,694,317	0

As a preview in 2019[6], UN Comtrade shows the alternative quantity alongside the quantity unit following the WCO recommendation (see Illustration 2). The reporting, processing and dissemination steps for quantity and weight information in UN Comtrade can be summarized as shown in Illustration 3.

Illustration 2: Example for the dissemination of trade data in UN Comtrade+ (with alternate quantity)[7]

Reporter	Partner	Commodity Code	Trade Value (US$)	Netweight (kg)	Qty Unit	Qty	Alternte quantity unit	Alternate quantity	Flag
Ireland	United Kingdom	854810	$675,035	252,678	Weight in kilograms	252,678	Number of cells/elem	1390	0

Table 1: WCO Supplementary Quantity Units

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**	Thousands of items
10	**U (jeu/pack)**	Number of packages
11	**12u**	Dozens of items
12	**m³**	Volume in cubic meters
13	**carat**	Weight in carats

[5] The columns Period (2018), Reporter (USA) and Partner (World) are not shown.
[6] https://comtrade.un.org/data (select preview)
[7] The columns Period (2017), Trade Flow (Export), 2nd Partner (World), Customs desc (All customs procedure codes) and Mode of Transport (All modes of transport) are not shown.

Table 2. List of UNSD Standard Quantity Units

Abbreviation	Description	Abbreviation	Description
A	All	**kg**	Weight in kilograms
m²	Area in square meters	**1000 KG**	Weight in 1000 kilograms
BBL	Barrels	**l alc 100%**	Litre pure (100 %) alcohol - l alc. 100%
hive	Beehive	**head**	Head
12u	Dozens of items	**kg/net eda**	Kilogram drained net weight
kWH	Electrical energy in kilowatt-hours	**kg C5H14ClNO**	Kilogram of choline chloride
1000 kWh	Electrical energy in thousands of kilowatt-hours	**kg P2O5**	Kilogram of diphosphorus pentaoxide
km	Length in kilometers	**kg H2O2**	Kilogram of hydrogen peroxide
m	Length in meters	**kg met.am.**	Kilogram of methylamines
-	No quantity	**kg N**	Kilogram of nitrogen
u	Number of items	**kg KOH**	Kilogram of potassium hydroxide (caustic potash)
U (jeu/pack)	Number of packages	**kg K2O**	Kilogram of potassium oxide
2u	Number of pairs	**kg NaOH**	Kilogram of sodium hydroxide (caustic soda)
TJ	Terajoule (gross calorific value)	**kg 90% sdt**	Kilogram of substance 90 % dry
1000u	Thousands of items	**kg U**	Kilogram of uranium
m³	Volume in cubic meters	**ct/l**	Carrying capacity in tonnes
l	Volume in liters	**Bq**	becquerels
1000 m³	Volume in thousand cubic meters	**gi F/S**	Gram of fissile isotopes
1000 L	Volume in thousands of liters	**GRT**	Gross register ton
carat	Weight in carats	**GT**	Gross tonnage
g	Weight in grams	**ce/el**	Number of cells/elements

Illustration 3: Reporting, processing and dissemination of weight and quantity in UN Comtrade

Storage of quantity and weight information in two (four) separate fields

Usage of quantity units (WCO/UNSD) for individual commodities (6-digit sub-headings or tariff line level) for the supplementary quantity

i. Reported in the WCO recommended unit (RU)
ii. Reported in the UNSD standard unit (SU)
iii. Possible to convert the reported quantity to the WCO recommended or UNSD standard unit
iv. Reported in non-WCO RU and non-UNSD SU → candidate for UNSD SU

Reported by country

Standardization of quantities/quantity codes to the UNSD standard units

Detection of extreme values

a) Obvious outliers → Countries are contacted
b) Detected by standard unit checks → Cell is blanked out and treated as missing
c) Report is created to measure share of outliers → To ensure no mistake in standardization of quantity units

Conversion and/or estimation if weight and quantity is not reported in WCO recommended unit, not reported or blanked out

a) Conversion (e.g., FAO conversion factors for selected commodities, i.e., liter to kg)
b) Estimation using (partially) reported unit values
c) Estimation using standard unit values
    1. High reliability SUV → estimated data to be shown at sub-heading level
    2. Low reliability SUV → estimated data is hidden at sub-heading level but values are aggregated to heading level

Processed by UNSD

Documentation of estimated values

Estimated records are

- identifiable by an estimation flag/code (0 = no estimation, 2 = quantity estimation only, 4 = net weight estimation only, 6 both quantity and net weight estimation), and
- are shown in italics on the UN Comtrade result page.

Estimated quantities with limited reliability are clearly identified by a quality flag.

Dissemination of estimated quantity and weight

Disseminated by UNSD

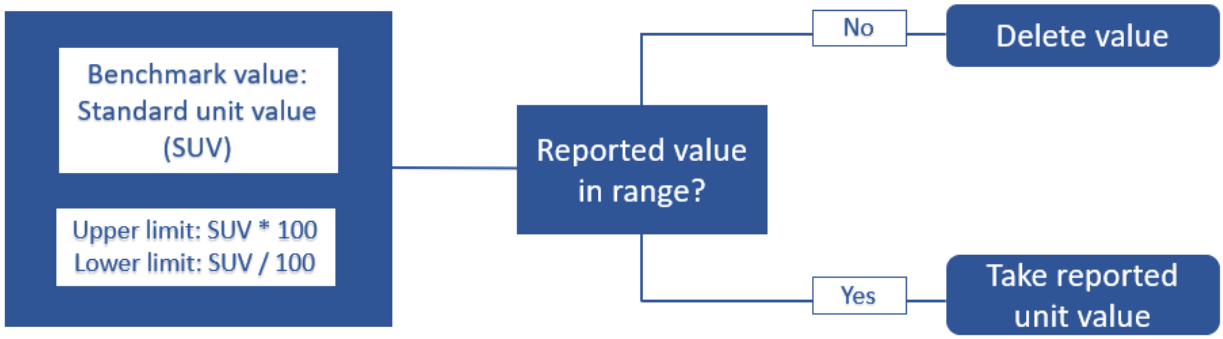
	High reliability of estimate	Low reliability of estimate
Used for aggregation	X	X
Disseminated	X	

I Outliers detection and correction methods used by UNSD

Reported quantity and weight values are deleted and tagged if the reported quantity or weight value is considered as an extreme value. In the further process, the "value" is treated as missing and is, where necessary and possible, estimated (see Section Estimation methods of missing data used by UNSD).

For the identification of extreme values, UNSD uses a standard unit value approach. The idea is to compare the reported unit value of commodity flows (6-digit exports and imports) aggregated to partner country World with a benchmark value (see also Illustration 4). If the reported value is in a specific range, the value is reported. Otherwise, the value is deleted and treated as missing.

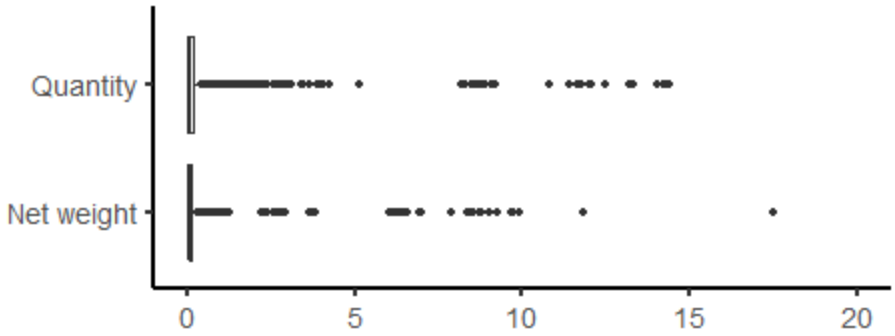
Illustration 4: Extreme value detection scheme



Note: For the calculation of standard unit values, see the Report on Standard Unit Values 2007 https://unstats.un.org/unsd/tradekb/Knowledgebase/50146/Quantity-and-Weight-Data-in-UN-Comtrade

The multiplication and division by 100 to obtain the upper and lower limit is a conservative definition of an extreme value. This is also expressed in the predominantly low shares of deleted extreme values. The records of processed data since the beginning of 2018 show that the median of the percentage of values that is identified as extreme value in reported data is only 0.099% for quantities and 0.094% for net weight. However, the boxplots in Figure 1 show that the spread is large due to some processed data sets, e.g., monthly data in 2017 for Mozambique, Indonesia and Slovenia.

Figure 1: Boxplot on share of extreme weight and quantity records to corresponding reported records in %[8]



Standard unit values (SUV)

Standard unit values are calculated

- at the end of the year,
- for all commodity flows (6-digit imports and exports),
- using the available data* of the latest reporting year.

The SUV of a specific commodity flow is defined as the **median unit value of all reporting countries from the prior year after the elimination of outliers**. SUVs are used for quantity and/or weight estimation and detection of outliers in the data of the subsequent reporting year.

*The calculation which was conducted in early 2019 had total records of 300 million tariff line level data.

II Estimation methods of missing data used by UNSD

UNSD aims to standardize country data for dissemination. Thus, quantity and weight information that is missing (which includes deleted extreme values) or not reported in the WCO recommended unit is either modified by conversion or estimated (see also Illustration 5), if possible.

Illustration 5: Unit value checks for reported data

The conversion factors used by UNSD are based on mathematical relations or suggestions of the Food and Agricultural Organization (FAO). The latter differ across commodity codes. The new quantity value can be derived using conversion factors as follows:

[8] 22 records (0.04% of all records) are not shown in Figure 1 due to the cut-off at 20%.

Converted quantity = Old quantity * Conversion factor.

The conversion is only based on known conversion factors and not on other reported values.

Example: Conversion

Table 2 shows the commodity code and description as well as the quantity of the commodity "cigarettes containing tobacco". The last two columns contain the reported unit code (unit code 5 equals numbers) and the recommended unit WCO (unit code 8 equals kg). Thus, the commodity is not reported as recommended.

Table 2: Original data for HS code 240220

HS Code	Description	Quantity	Unit	WCO RU
240220	Cigarettes containing tobacco	4000000	5	8

In this example, the quantity unit 5 must be converted to quantity unit 8. According to FAO conversion factors (see e.g., International Merchandise Trade Statistics: Compilers Manual)[9] the conversion factor for the HS code 240220 is 0.001. The converted quantity can easily be derived following the equation above.

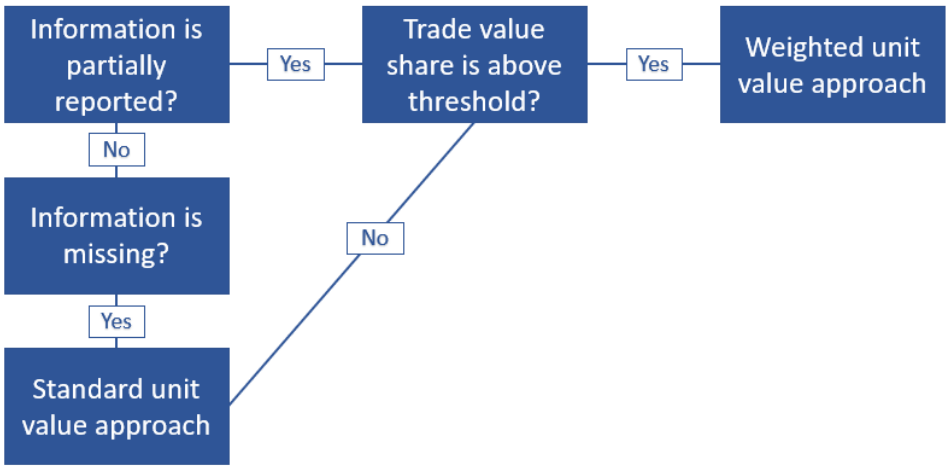
Table 3: Original data for HS code 240220

HS Code	Description	Quantity	Unit	WCO RU
240220	Cigarettes containing tobacco	4000000 * 0.001 4000	8	8

Estimation approaches are used when a conversion factor is not available, or the value is missing. UNSD uses two different approaches: i) a weighted unit value, and ii) a standard unit value approach.

[9] Department of Economic and Social Affairs, Statistics Division (2013), International Merchandise and Trade Statistics: Compilers Manual, Revision 1 (IMTS 2010-CM), Annex XV.B. https://unstats.un.org/unsd/trade/EG-IMTS/IMTS2010-CM%20-%20white%20cover%20version.pdf

Illustration 6: Decision process for the estimation method for each 6-digit commodity code



For the **weighted unit value approach**, reported or rather partially reported quantity and weight information is used to estimate non-reported quantity and weight within the same commodity flow (6-digit imports and exports) of a country. A weighted unit value derived from the value/quantity or value/weight ratio of the properly reported data is used to estimate the missing quantity and weight information:

Estimated Quantity = Value / Partially Reported Unit Value

This approach is assumed to be reliable if the trade value share of the data with properly reported quantity and weight passes a certain threshold, which is 20% for OECD and 50% for non-OECD countries. Otherwise, the estimated value is considered as not reliable and the standard unit value approach is used (see below). Quantities are estimated at the most detailed level of reported data. In cases where countries apply different quantity units within the same 6-digit commodity heading, the estimation is attempted first at tariff line level for the calculation using weighted unit value, and then the units are converted into standard units for aggregation.

Example: Weighted unit approach

Table 4 shows the commodity code and description as well as the value and the quantity of the commodity "fresh apple". The last two columns contain the reported unit code (unit code 5 equals numbers) and the recommended unit (RU) code by the WCO (unit code 8 equals kg). Thus, the commodity is partly reported as recommended. However, for one part the unit is numbers and not as recommended. Furthermore, there is no conversion factor for HS code 080810 from number to kg. Therefore, the unit value needs to be estimated.

Table 4: Original data

HS Code	Description	Value	Quantity	Unit	WCO RU
080810	Fresh apple	200	80	5	8
080810	Fresh apple	400	100	8	8

The value share and the unit value can be calculated from the value and the quantity.

Value share = Value / Total Value
Unit value = Value / Quantity

Thus, the following value shares and unit values can be derived.

Table 5: Original data with calculated value share and unit value

HS Code	Description	Value	Value Share	Quantity	Unit	WCO RU	Unit value
080810	Fresh apple	200	200 / 600 33.33%	80	5	8	200 / 80 2.5
080810	Fresh apple	400	400 / 600 66.66%	100	8	8	400 / 100 4

Since the value share of the properly reported unit value is 66.6% and thus above both thresholds (20% for OECD member countries, 50% for non-OECD member countries), the new quantity can be estimated using the weighted unit value approach.

Table 6: Final data

HS Code	Description	Value	Quantity	Unit	WCO RU	Unit value
080810	Fresh apple	200	200 / 4 50	8	8	4
080810	Fresh apple	400	100	8	8	4

The **standard unit value approach** is used whenever the weighted unit value approach is not usable due to missing data or assumed to be unreliable, e.g., when the required threshold of the value share of the properly reported unit values is not met. A standard unit value is used to estimate the missing quantity and weight information:

Estimated quantity = Value / Standard Unit Value

This approach crucially depends on the information on previous reporting years.

Example: Standard unit value approach

Table 7 shows the commodity code and description as well as the value and the quantity for the commodity "fresh apple". The last two columns contain the reported unit code (unit code 5 equals numbers) and the recommended unit by the WCO (unit code 8 equals kg). The commodity is partly reported as recommended. In comparison to the previous example, the first value is now 1100 instead of 200.

Table 7: Original data

HS Code	Description	Value	Quantity	Unit	WCO RU
080810	Fresh apple	1100	80	5	8

080810	Fresh apple	400	100	8	8

The following value shares and unit values can be derived following the calculations in the previous example.

Table 8: Original data with calculated value share and unit value

HS Code	Description	Value	Value Share	Quantity	Unit	WCO RU	Unit value
080810	Fresh apple	1100	1100 / 1500 73.33%	80	5	8	1100 / 80 13.75
080810	Fresh apple	400	400 / 1500 26.66%	100	8	8	400 / 100 4

In contrast to the previous example, it makes a difference if the country is an OECD country or not. In the latter case, the weighted unit value approach should not be used since the value share of properly reported data is only 26.66% and thus below the threshold of 50%. For OECD countries, the weighted unit value approach as shown in the previous example can still be used. The standard unit approach is based on information of previous years. Table 6 shows the standard unit values for the HS code 080810 of the last reported year.

Table 9: Standard unit value for HS code 080810 for the last reported year

HS Code	Description	Flow	Unit	SUV	Low unit value	High unit value
080810	Fresh apple	Import	8	0.898	0.00898	89.8
080810	Fresh apple	Export	8	0.249	0.00249	24.9

Table 10 shows the estimated quantity assuming the original data in Table 8 is import data.

Table 10: Final data

HS Code	Description	Value	Quantity	Unit	WCO RU	Unit value
080810	Fresh apple	1100	1100 / 0.898 1,225	8	8	0.898
080810	Fresh apple	400	100	8	8	4

III Aggregation methods by UNSD

For users' convenience, some data dimensions are pre-aggregated so that the values can be extracted without performing calculations first. In UN Comtrade, the pre-aggregated data includes the trading partner "world", total exports and imports, higher level commodities (total, HS: 2-, 4-, 6-digit levels, SITC: 1-, 2-, 3-, parts of 4-digit levels, and BEC). Trade values and quantities are aggregated. The aggregation levels for the trade values that are always disseminated in US$ are shown in Table 11.

Table 11: Aggregation levels for trade values

Aggregation level	Explanation
Partner countries	Partner "world" is the sum of individual partner countries (or statistical territories); should be applied to 1st and 2nd partner countries.
Trade flows	Total imports and exports consist of identified sub-flows (i.e., Re-exports).
Products	Detailed commodity categories are added to the higher level of aggregation: if reported data of the higher level exists, the reported data is kept.
Customs procedure codes	"All" customs procedures is the sum of all other sub customs procedure codes.
Mode of transport	Aggregation only at highest level of hierarchy, "all modes of transport", even though different levels are available.

For the aggregation of quantity information (net weight and quantity) some rules need to be considered:

1. Quantity information can only be aggregated up to the 4-digit level in HS, the 3-digit level in SITC and there is no quantity information in BEC.
2. It can only be aggregated if the value of quantity is not zero because this would impact the result (of the unit value) at higher level of aggregation (i.e., partner World or 4-digit HS).
3. It can only be aggregated if all underlying records have the same quantity unit; otherwise it would be set to "no quantity unit"
4. If underlying records contain an estimation flag, the estimation flag is carried over to resulting aggregated record.
5. The visibility flag (of net weight or quantity) is always set to 1 for pre-aggregated data.

IV Proposed improvements for UN Comtrade/Self-assessment

Creation of SUV

- SUV was developed for combination of periods, sub-headings, quantity units and flows. For better granularity should it be expanded to regionals (of reporters and/or partners) and mode of transports (i.e., is the transport by ship cheaper than by plane)?
- Should we build standard unit values for different country groups instead of World?

Outliers' detection

- Outliers' detection is based on SUV with arbitrary multiplication of 100. This leads to 0.1% extreme unit value detection in country datasets. Should we review this constant?

Standard Units

- The standard units of quantity are set for HS sub-headings, and it is important that we have standard units for HS headings. This would allow proper estimation and aggregation at heading level
- UNSD has introduced alternate quantity units that extend WCO RU; however, should we set a standard for sub-heading?

Estimation methodology

- Are there any updated FAO conversion factors? Should we supplement it with other conversion factors such as converting petroleum from volume to weight or vice-versa?
- Should we review different threshold between OECD and non-OECD countries? Or the categorization criteria of country groups?
- UNSD estimation methodology does not take into account historical series; should we implement historical regression?
- Current methodology uses traditional statistical methods, should we explore more on machine learning approach?
- The methods of Weighted Unit Value and Standard Unit Value are presently applied to 6 digits level (whether or not country reports data at tariff line level). The accuracy can be improved by applying those methods to tariff line level (considering that there are enough data points).