Integration of simple data validation tools into the UNECE production tables

Michael Nagy

13th Session of the Joint Task Force on environmental Statistics and Indicators, 29-30 June 2017, Geneva

UNECE

Content

1. Why this presentation
2. What is data validation?
3. Are there internationally agreed processes for data validation?
4. Data validation examples from other International Organisations
5. Recommendations for the UNECE production templates

1. Why this presentation

- Decision taken at the 11th Session of the JTF
- See following paragraph 31 of the report:

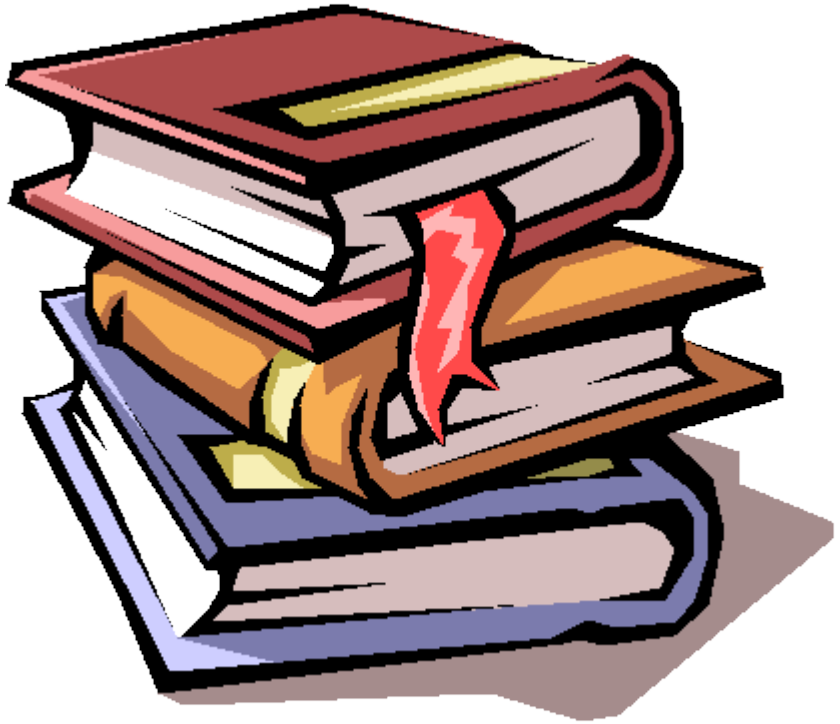
31. The need for specific guidelines for quality assurance (QA) of environment statistics and metadata as part of the SEIS self-assessment was highlighted. The Joint Task Force called for further action towards producing such guidelines, which should complement existing QA frameworks used in countries and by international bodies. Countries asked for case studies and examples of good practices in quality assurance of environment statistics. The integration of simple validation tools into the UNECE production table (similar to validation tools already used by Eurostat and UNSD) could be a first practical step to support quality assurance.

2. What is data validation?

Definition of data validation

An activity aimed at verifying whether the value of a data item comes from the given set of acceptable values. (UNECE glossary on statistical data editing, Eurostat and OECD use similar definitions)

3. Are there internationally agreed processes for data validation?

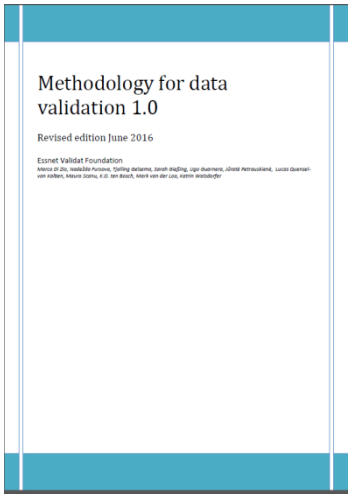


Procedures and approaches for data validation have never been systematised

Data validation is carried out by all NSOs, but:

- procedures and approaches have never been systematised
- most of them are ad-hoc.

In June 2016 Eurostat has published a first document “Methodology for data validation 1.0” as an output of the Essnet project Validat foundation, primarily financed by Eurostat, involving CBS, Destatis, Istat and Statistics Lithuania.
https://ec.europa.eu/eurostat/cros/system/files/methodology_for_data_validation_v1.0_rev-2016-06_final.pdf

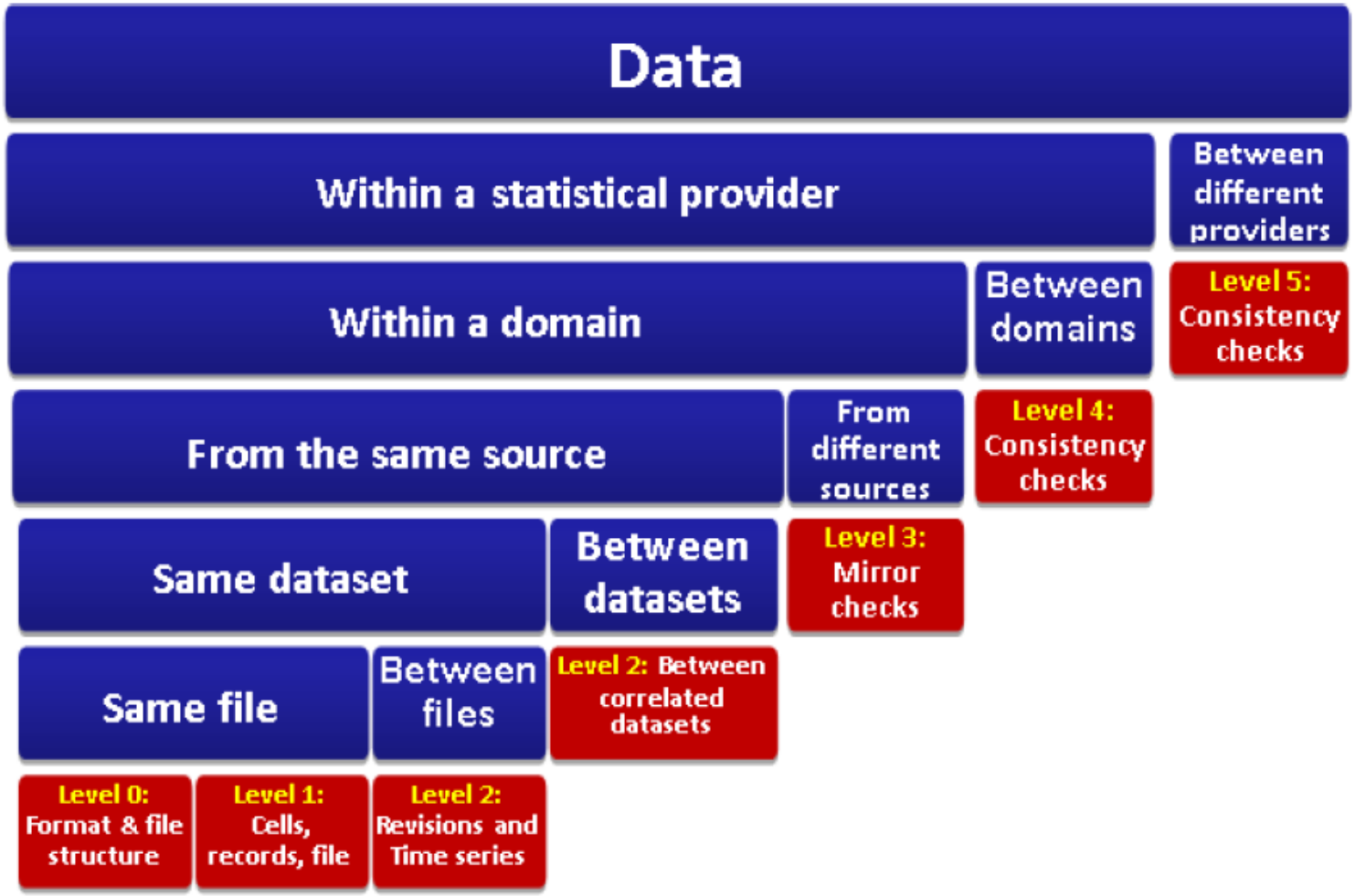

Methodology for data validation 1.0

Revised edition June 2016

Essnet Validat Foundation

Eurostat methodology for data validation 1.0

Data validation levels



Eurostat methodology for data validation 1.0

Validation rules can be categorised as follows:

A – integrity of the data file

- Formal validity of entries
- Presence of an entry
- No duplicate entries
- Code lists (if applicable) are used

B - Logical validation and consistency

Table 1: Categories of a 2-way typology for validation rules for logical validation and consistency

Typology dimension	Types of checks	
1	Identity checks	Range checks • bounds fixed • bounds depending on entries in other fields
2	Simple checks, based directly on the entry of a target field	More "complex" checks, combining more than one field by functions (like sums, differences, ratios)

+ conditional checks

Eurostat methodology for data validation 1.0

Examples for logical validation and consistency

- **Conditional check:** If a country is landlocked, then the discharge to the sea has to be 0
- **Identity check:** The total water use has to be equal to water available for use minus losses
- **Range check with fixed bounds:** Water use of households per capita should be between 100 l/day and 400 l/day
- **Range check with bounds depending on entries in other fields:** water losses must me smaller than water available for use

4. Data validation examples from other International Organisations

UNSD/UNEP Environment Questionnaire

• If the value turns red, please check if it is correct.

DATA VALIDATION

Values not validated against checks like those below cannot be included in the UNSD Envstats Database and dissemination products unless sufficient explanation is provided in footnotes.

If footnotes explanations are already provided, the relevant validation alerts can be ignored.

Section: WASTE

Country: Contact institution:

Table R1: Generation of Waste by Source

• Unhide to view/edit previous years • If the value turns red, please check if it is correct.

Line	Category	Unit	1990	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
1	Agriculture, forestry and fishing (ISIC 01-03)	1000 t													
2	Mining and quarrying (ISIC 05-09)	1000 t													
3	Manufacturing (ISIC 10-33)	1000 t													
4	Electricity, gas, steam and air conditioning supply (ISIC 35)	1000 t													
5	Construction (ISIC 41-43)	1000 t													
6	Other economic activities excluding ISIC 38	1000 t													
7	Households	1000 t													
8	**Total waste generation (=1+2+3+4+5+6+7)**	1000 t													

Page 1

Notes:

- Waste generated by an economic activity includes all kinds of waste generated by economic units within this activity. For further details please refer to the description of Table R1 in the Guidance Section.
- **If the requested data are not available, please leave the cell blank. If the requested variable is not applicable (the phenomenon is not relevant) to the country or the value is less than half the unit of measurement, the cell should be filled with "0".**
- Please provide in the Footnotes Section below information on the source and data collection methodology for the values provided, such as estimation methods (if any), and the types of the original data sources used (e.g., surveys or administrative records).
- **Data can also be viewed/edited for years 1995-2003. Select column G to column Z, right-click, and select "Unhide".**

Footnotes

DATA VALIDATION

Values not validated against checks like those below cannot be included in the UNSD Envstats Database and dissemination products unless sufficient explanation is provided in footnotes.

If footnotes explanations are already provided, the relevant validation alerts can be ignored.

Time series validation: an automatic check on the percentage change from the previous year (where available). Cases are flagged if changes are not within expected ranges.

Line	Category	Unit	1990	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
1	Agriculture, forestry and fishing (ISIC 01-03)	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	Mining and quarrying (ISIC 05-09)	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	Manufacturing (ISIC 10-33)	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	Electricity, gas, steam and air conditioning supply (ISIC 35)	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5	Construction (ISIC 41-43)	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6	Other economic activities excluding ISIC 38	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
7	Households	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
8	**Total waste generation (=1+2+3+4+5+6+7)**	1000 t	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Coherence validation: check of coherence between variables within the questionnaire or compared to data from outside sources. Cases are flagged if values are not within expected ran

Line	Category	Unit	1990	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
8	**Total waste generation (=1+2+3+4+5+6+7)**	1000 t	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	**Line 1+2+3+4+5+6+7**	1000 t	0	0	0	0	0	0	0	0	0	0	0	0	0	0
√	**Line 8= 9**		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
10	**Total waste per $1000 GDP**	t/$1000	####	####	####	####	####	####	####	####	####	####	####	####	####	####
√	**.05<Total waste per $1000 GDP<10**		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

UNSD/UNEP Environment Questionnaire

Automatic validation by UNSD:

- Coherence:
  - E.g. total waste generation is not equal to the sum of waste generated by economic activities and households – value turns red (note: on purpose the questionnaires of UNSD do not calculate sums automatically)
  - Comparison with data from outside sources (e.g. World Bank, FAO Aquastat etc.)
- Time series
  - Significant changes will turn red in the grey data validation template (e.g. total amount of municipal waste collected increases by >25% within one year).

International Energy Agency
Monthly Oil Questionnaire

MONTHLY OIL QUESTIONNAIRE

Country	*(select country)*
Month of data	*January* *2013*
Status of data (final/provisional)	Final
Prepared by	
Date of transmission	

ALL DATA SHOULD BE ENTERED
IN THOUSAND METRIC TONS
WITH NO DECIMAL PLACE

Table 1: SUPPLY OF CRUDE OIL, NGL, REFINERY FEEDSTOCKS, ADDITIVES AND OTHER HYDROCARBONS

Unit: Thousand metric tons

			Crude oil	Natural gas liquids	Refinery feedstocks	Additives / oxygenates	*Of which Biofuels*	Other hydrocarbons	**Total (A to F, excl. E)**	**Probate**
			A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
+	Indigenous production	1	10						10	
+	Receipts from other sources	2							0	
+	Backflows[1]	3							0	
+	Products transferred[2]	4							0	
+	Imports (Balance)[3]	5	50						50	
-	Exports (Balance)[4]	6	80						80	
-	Direct use[5]	7	1						1	
-	Stock changes[6]	8	0						0	
=	Refinery intake (Calculated)	9	-21	0	0	0	0	0	-21	
-	Statistical difference	10	-46	0	0	0	0	0	-46	
=	***Refinery intake (Observed)***	11	***25***						***25***	
	Memo item: Refinery losses	12							0	

1. Total of Backflows (reference G3) should correspond to Total products of Backflows to refineries in Table 2 (reference Z19). — OK
2. Total of Products transferred (reference G4) should correspond to Total products of Products transferred in Table 2 (reference Z9). — OK
3. Imports (Balance) should correspond to Total imports for these products in Table 3 (references A105 to F105). — Total Imports (Balance) do not equal Total Imports in Table 3
4. Exports (Balance) should correspond to Total exports for these products in Table 4 (references A96 to F96). — Total Exports do not equal Total Exports in Table 4
5. Direct use should be carried over to Primary product receipts in Table 2 (line 1). — Total Direct use figure does not equal Total products of Primary product receipts in Table 2
6. Stock changes should correspond to Closing level minus Opening level of Stocks on national territory, Table 5 (line 2 minus line 1). — OK

3. Recommendations for the UNECE production templates

General recommendations for automatised validation rules in the UNECE data production tables

Main goals of automatised validation:

- Identify wrong use of units of measurement
- Identify typos (e.g. wrong decimals)
- Identify unexpected changes in time series
- Identify if data are balanced (when there should be a balance)

Validation should indicate unexpected values, but

- Not block data entry
- Invite data providers to double-check, correct if necessary, and provide more information in footnotes

Validation will be based on data within the same file

Logical validation and consistency checks only

Test with energy indicators and biodiversity indicators, then gradual implementation

Side condition

No formulas for automatic calculation of sums (e.g. total water use is calculated automatically as the sum of water use by economic activities + households)

- Sometimes only the total is known, but not the disaggregated values.
- A sum would be calculated automatically, even if not all disaggregated values are available.
- Automatic formulas for the data entry would outsmart parts of the data validation

Recommended validation rules to be implemented into the UNECE data production templates

Simple conditional checks:

- e.g. landlocked countries cannot have discharge to the sea or marine protected areas

Identity checks:

- E.g. the total primary energy supply be equal to production + imports – exports – bunkers + stock changes
- E.g. the total waste generated has to be the sum of waste generated by economic activities + households

Range check with fixed bounds

- E.g. Water use of households per capita should be between 100 l/day and 400 l/day

Range check with bounds depending on entries in other fields:

- e.g. tolerance values for changes over time
- e.g. water losses must me smaller than water available for use

3. Questions, comments?

- Does the JTF agree with the recommended validation rules?

- Should the secretariat test it with the revised templates for energy and biodiversity indicators?

- Any other comments?