Waterbase – Rivers
Version 10

Quality control documentation

31 May 2010

Waterbase – Rivers

Data on rivers are collected annually through the WISE-SoE data collection process. Data and information obtained through the WISE-SoE data collection process are primarily used to compile indicator factsheets, associated with the EEA's Core Set Indicators, upon which EEA assessment reports are based. Data collected through the WISE-SoE data collection process are also published in WISE map viewer, Waterbase, a series of water topic-specific databases and web pages, publicly accessible via the EEA Data Service's web site.

The dataset contains data on nutrients and organic matter, proxy pressure data on the upstream catchments and physical characteristics of the WISE-SoE river monitoring stations.

QA/QC activities

This document briefly presents the ETC/Water and the EEA activities focused on quality of the Waterbase - Rivers dataset and the results of these activities. In addition a warning is given on the use of certain records for analytical purposes (see section 2, 3 and 4).
The Quality control tests have been performed on the Waterbase - Rivers database provided in April 2010 by ETC/WTR. This database is included in the EEA data service as version 10, and is publicly available. The database and metadata are available at the following URL: http://www.eea.europa.eu/data-and-maps/data/waterbase-rivers-6

Waterbase – Rivers dataset contains three data tables:

- QUALITY
- STATIONS
- PRESSURES

Five types of the tests have been performed on the data tables. Basic tests, Logical rules violation test, Outlier detection, Stations tests, and Valid data type and codes tests.

Chemical rules tests, that were first time introduced previous year, were also tested by the ETC/Water but the result were not incorporated due to critical comments obtained from countries as a reaction to number of false positive errors detected by the tests. The rules are presently being revised.

1. Basic tests

1.1 Summary

1.1.1 Waterbase - Rivers: QUALITY

Country	Number of records								
	from the last delivery					in the ETC working database		Waterbase	
	total	inserted into working database							
		new		redelivered old					
		total	QA issue	total	QA issue	total	QA issue	**total**	QA issue
AL	510	510	69	0	0	3037	416	3037	66
AT	777	777	0	0	0	26051	1800	26051	138
BA	543	543	3	0	0	3513	514	3513	15
BE	788	788	3	0	0	8116	1557	8116	369
BG	1086	1086	0	0	0	12291	565	12291	8
CH	815	97	0	718	0	1852	210	1852	
CY	224	224	0	0	0	578	72	578	18
CZ	851	851	2	0	0	13750	2254	13750	43
DE	2536	2536	13	0	0	28636	7770	28636	3635
DK	415	207	0	218	0	12853	26	12853	5
EE	900	900	0	0	0	12550	1769	12550	2
ES	16690	16690	41	0	0	87448	14901	86293	6479
FI	5781	5781	1	0	0	162124	2418	161726	160
FR	28489	28489	1	0	0	158547	14287	158547	26
GB	4160	1168	0	2992	0	34750	1564	34750	793
GR	0	0	0	0	0	2093	322	2093	194
HR	769	769	0	0	0	6839	1823	6839	
HU	10600	0	0	547	522	94587	21675	94040	7523
IE	579	579	0	0	0	6022	501	5833	239
IS	14	14	0	0	0	74	13	74	
IT	9826	9826	44	0	0	32915	3001	32913	353
LI	8	4	0	4	0	16	0	16	2
LT	800	800	0	0	0	21841	2968	21788	10
LU	26	26	0	0	0	354	0	354	4
LV	600	600	1	0	0	11235	1148	11235	
ME	0	0	0	0	0	0	0		
MK	177	177	182	0	0	2421	420	2421	137
NL	348	172	0	176	0	4701	797	4701	5
NO	2185	368	0	0	0	2785	402	2785	
PL	1463	1463	0	0	0	27619	5855	27619	410
PT	328	328	0	0	0	1452	206	1430	83
RO	1401	1401	1	0	0	7621	37	7621	17
RS	1407	1407	2	0	0	6127	926	6127	2
SE	1295	1295	0	0	0	39802	91	39802	88
SI	139	139	0	0	0	4854	839	4854	20
SK	1460	1460	0	0	0	9337	1177	9337	511
TR	120	30	2	0	0	120	13	120	1
Total	98110	81505	365	4655	522	848911	92337	846545	21356

1.1.2 Waterbase - Rivers: STATIONS

Country	Number of records								
	from the last delivery					in the ETC working database		Waterbase	
	total	inserted into working database							
		new		redelivered old					
		total	QA issue	total	QA issue	total	QA issue	**total**	QA issue
AL	51	0	0	51	0	52	0	52	
AT	71	0	0	71	0	290	0	290	
BA	46	4	0	42	0	56	0	56	
BE	63	5	0	58	0	67	1	67	4
BG	98	0	0	98	0	111	0	111	
CH	21	14	0	7	0	22	0	22	
CY	31	0	0	31	0	33	0	33	
CZ	0	0	0	0	0	73	0	73	
DE	260	116	0	144	0	267	1	267	31
DK	42	0	0	42	0	42	0	42	
EE	61	0	0	61	0	61	0	61	
ES	1405	1315	0	90	0	2829	0	2829	
FI	138	0	0	138	0	230	0	230	
FR	1524	8	0	1516	0	1947	0	1947	
GB	177	2	0	175	0	206	0	206	
GR	0	0	0	0	0	94	5	94	94
HR	45	5	0	40	0	50	0	50	
HU	97	0	0	97	0	152	0	152	96
IE	180	0	0	180	0	209	0	209	
IS	2	1	0	1	0	2	0	2	
IT	1004	88	0	916	0	1490	0	1490	107
LI	1	1	0	0	0	1	0	1	
LT	58	4	0	54	0	102	0	102	
LU	4	0	0	4	0	4	0	4	
LV	43	8	0	35	0	118	0	118	
ME	0	0	0	0	0	0	0		
MK	20	0	0	20	0	20	1	20	1
NL	16	8	0	8	0	31	0	31	
NO	46	0	0	46	0	46	0	46	
PL	134	0	0	134	0	136	0	136	
PT	59	0	0	59	0	59	4	59	4
RO	118	0	0	118	0	126	0	126	
RS	76	0	0	76	0	77	0	77	
SE	122	0	0	122	0	127	0	127	
SI	0	0	0	0	0	30	0	30	
SK	89	8	0	81	0	131	0	131	95
TR	5	0	0	5	0	5	0	5	
Total	6107	1587	0	4520	0	9296	12	9296	432

1.1.3 Waterbase - Rivers: PRESSURES

Country	Number of records								
	from the last delivery					in the ETC working database		Waterbase	
	total	inserted into ETC working database							
		new		redelivered old					
		total	QA issue	total	QA issue	total	QA issue	**total**	QA issue
AL	51	7	0	44	0	52	0	51	21
AT	71	0	0	71	0	292	2	141	
BA	0	0	0	0	0	0	0		
BE	0	0	0	0	0	59	1	59	4
BG	0	0	0	0	0	0	0		
CH	21	14	0	7	0	22	0	22	22
CY	31	0	0	31	0	32	0	32	
CZ	0	0	0	0	0	72	0	72	
DE	242	116	0	126	0	263	1	243	6
DK	0	0	0	0	0	42	0	42	
EE	61	0	0	61	0	61	0	61	
ES	1405	1315	0	90	0	2829	0	2829	
FI	0	0	0	0	0	5	0	1	
FR	1524	8	0	1516	0	1885	0	1876	
GB	0	0	0	0	0	190	0	190	
GR	0	0	0	0	0	0	0		
HR	0	0	0	0	0	0	0		
HU	97	0	0	97	0	152	0	151	2
IE	180	0	0	180	0	209	0	180	180
IS	2	1	0	1	0	2	0	2	1
IT	57	57	0	0	0	57	0	57	
LI	1	0	0	1	0	1	0	1	
LT	58	4	0	54	0	102	0	102	
LU	4	0	0	4	0	4	0	3	
LV	0	0	0	0	0	110	0	109	
ME	0	0	0	0	0	0	0		
MK	0	0	0	0	0	20	1		
NL	0	0	0	0	0	12	0		
NO	46	0	0	46	0	46	0	46	
PL	134	0	0	134	0	136	0	136	
PT	59	0	0	59	0	59	4	59	4
RO	0	0	0	0	0	124	0	1	
RS	0	0	0	0	0	0	0		
SE	122	0	0	122	0	127	0	127	
SI	0	0	0	0	0	24	0	24	
SK	86	6	0	80	0	129	0	129	7
TR	0	0	0	0	0	0	0		
Total	4252	1528	0	2724	0	7118	9	6746	247

1.2 Primary key tests

Primary key is a field or combination of fields with values which have to be unique in the data table. If primary key is duplicated it is an error which has to be solved.

List of data tables primary keys:

- QUALITY: CountryCode, Waterbase_ID, Determinand, Year, AggregationPeriod
- STATIONS: CountryCode, Waterbase_ID
- PRESSURES: CountryCode, Waterbase_ID

Result:
No primary key error has been detected.

1.3 Table relations tests

The unique Waterbase identifier (WaterbaseID) is present in each of the data tables. It can be used to link data from one table to another. The table relations tests detect identifiers which are not present in some of the tables.

1.3.1 Number of Stations without any data in the "QUALITY" table by country

Country Code	No. of stations	Percentage of total no. of ststions
BA	6	10.71%
CY	10	30.30%
DE	4	1.50%
EE	1	1.64%
ES	702	24.81%
FR	118	6.06%
GR	9	9.57%
HU	2	1.32%
IE	10	4.78%
IT	12	0.81%
LT	4	3.92%
LV	1	0.85%
PT	3	5.08%
SK	9	6.87%
Total	891	9.58%

1.3.2 Number of Stations without any data in the "PRESSURES" table by country

Country Code	No. of GW bodies	Percentage of total no. of GW bodies
AL	1	1.92%
AT	151	52.07%
BA	56	100.00%
BE	8	11.94%
BG	111	100.00%
CY	1	3.03%
CZ	1	1.37%
DE	24	8.99%
FI	229	99.57%
FR	71	3.65%
GB	16	7.77%
GR	94	100.00%
HR	50	100.00%
HU	1	0.66%
IE	29	13.88%
IT	1433	96.17%
LU	1	25.00%
LV	9	7.63%
MK	20	100.00%
NL	31	100.00%
NO	3	6.52%
RO	125	99.21%
RS	77	100.00%
SI	6	20.00%
SK	2	1.53%
TR	5	100.00%
Total	2555	27.48%

1.3.3 “QUALITY” and “PRESSURES” table records where none of the stations is present in the “STATIONS” table

Table	Country Code	No of records	Percentage of total no of records
QUALITY	HU	3	0.003%
QUALITY	RO	1	0.013%
QUALITY	SK	1	0.011%
QUALITY	Total	5	0.001%
PRESSURES	AT	2	1.418%
PRESSURES	Total	2	0.030%

All of these records are marked in the dataset (see section 4 for more details)

2. Logical rule violation tests

Logical rules were tested in the “QUALITY” data table. This table contains several measurement value fields, calculated in the aggregation process. Logical relations can be detected between them and mathematically transformed in a set of rules. Following rules have been detected and tested:

Rule	Basic validation rules
1	Mean >= Minimum
2	Mean <= Maximum
3	Median >= Minimum
4	Median <= Maximum
5	Minimum <= Maximum
6	StandardDeviation < Maximum
Rule	**Combined validation rules**
13	IF Minimum < Maximum THEN (StandardDeviation > 0)
14	IF NumberOfSamples = 1 THEN (Mean = Minimum = Maximum = Median)
15	IF NumberOfSamples = 1 THEN (StandardDeviation = 0)
16	IF NumberOfSamples = 0 THEN (AllValueType Is Null)
Rule	**Negative value validation rule**
17	All Values >=0

The following exceptions and modifications were been applied:

IF Maximum = 0 AND StandardDeviation = 0 THEN rule 6 is not violated

A special QA field (QA_LRviolations) has been added to the data tables. Information of the rules violated in the respective record are kept there as a coma separated list of those rules numbers (the numbers are the same as in the table above). It is recommended that the records where QA_LRviolation field is not empty **(2124 Quality records)**, should not be used in a further analysis or only after a careful consideration. The detected data quality inconsistencies will be tried to be solved in the near future.

3. Outlier detection

Detection of outliers was performed on the “QUALITY” data.

Measurement “Mean” values were tested against limiting values individually defined by an expert for each of the determinands and also statistically compared with other values from the same time series. If the value was detected as an outlier it was analyzed whether it can be a possible error or whether it was caused by natural conditions.

Records where Mean value is not provided are also acknowledged as outliers.

The findings described above have been stored in a special QA field (QA_outlier) added to data table. Following QA flags have been used:

-2 – measurement has been confirmed to be taken from a highly polluted area **(80 Quality records)**

-1 – record has been confirmed by the respective country as being correct **(152 Quality records)**

1 – standard potential outlier - value is either higher/lower than limit value or is suspiciously high/low comparing to the rest of the time series or value change between two consecutive values is suspiciously abrupt or was marked as an potential outlier by a content expert **(713 Quality record)**

2 – measurements are probably taken from a highly polluted locations **(124 Quality records)**. It is recommended not to use them for calculation of average concentrations of nutrients for broader areas like RBD or whole Country.

3 – the whole or a big part of the particular country delivery is considered as problematic because it contains too many quality issues **(5959 Quality records: 5940 records HU 2007, 19 records MK 2006)**

10 – the Mean value = 0 **(7337 Quality records)**. Value is not correct and records should not be used.

99 – the Mean value is empty **(4336 Quality records)**. Record can’t be used.

4. Stations tests

Positions of all reported monitoring stations have been tested using the coordinates provided as well as stations availability. The cases when the station coordinates fall outside the respective country borders, when coordinates are missing or when the monitoring station available in the Quality or Pressures data tables is not available in the Stations table, are documented in a special QA field (QA_station_problem). In addition some other station related issues were tested. Following QA flags have been used:

1 – monitoring station is located outside the respective country borders – either on the sea or in another country **(1 station, 102 Quality records)**

2 – coordinates are missing **(5 stations, 67 quality records)**

4 – more stations with the same coordinates **(21 stations, 456 quality records, 7 Pressures records)**

5 – water category does not belong to Rivers **(6 stations, 478 quality records, 6 Pressures records)**

99 – station is not available in the Stations table (**233 Quality records, 2 Pressures records)** – see result 1.3.3

These data quality inconsistencies will be tried to be solved in the near future.

5. Data type and codes tests

All Rivers dataset values have to follow specifications defined in the respective Data dictionary (DD) definitions. Values, which are of a different data type as requested (e.g. string instead of numeric) or which are not available in a set of allowable values, have been either removed or, if possible, replaced by a correct value.

There is one exception from this rule. Some of these “errors” are only formally wrong. The value is still valid but was not foreseen as possible and was therefore omitted to be included in the current DD definitions of the respective table field. In this case the original code has been left in the field untouched. It is planed that these codes will be added into the code list during the next DD update.

In all the cases the original, incorrect value or value missing in the DD code list, has been stored in a special QA field (QA_datatype_error) in the following format:

Name_of_field: Erroneous_Value; [Name_of_field: Erroneous_Value; …]

Test result summary:

Quality table: 582 records
Stations table: 411 records
Pressures table: 234 records

The cases where the errors couldn’t be corrected will be tried to be solved in the near future in cooperation with the data providers.