Waterbase – Rivers
Version 7

Quality control documentation

30 March 2007

Waterbase – Rivers

Data on rivers are collected annually through the Eionet-Water process. Data and information obtained through the Eionet-Water 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 Eionet-Water process are also published in Waterbase, a series of water topic-specific databases and web pages, publicly accessible via the EEA Data Service's web site.

Rivers dataset include physical characteristics of the river monitoring stations, proxy pressures on the upstream catchment areas, as well as chemical quality data on nutrients and organic matter in rivers.

QA/QC activities

This document briefly presents EEA activities focused on quality of Waterbase - Rivers dataset and results of these activities. In addition warning is given on the use of certain records for analytical purposes (see section 2 and 3).

Quality control tests have been performed on the Waterbase - Rivers database provided in December 2006 by ETC/WTR. This database is included in the EEA data service as version 7, and is publicly available. The database and metadata are available at the following URL: http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=984

Waterbase – Rivers dataset contains three data tables:

- PRESSURES
- QUALITY
- STATIONS

Three type of test have been performed on the data tables. Basic tests, Logical rules violation test and Outlier detection.

1. Basic tests

1.1 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.

List of data tables primary keys:
STATIONS: WaterbaseID
PRESSURES: WaterbaseID
QUALITY: WaterbaseID, Determinand, Year, AggregationPeriod

Result:
No primary key error has been detected.

1.2 Table relations tests

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

1.2.1 Number of "QUALITY" table records by country where WaterbaseID is not present in the "STATIONS" table

Country code	No. of records
TOTAL	0

1.2.2 Number of stations without any data in the "QUALITY" table by country

Country code	No. of stations	Percentage of total no. of stations
BA	6	14.63
ES	20	4.67
FR	7	1.25
GR	9	9.57
SK	3	5
TOTAL	45	1.25

1.2.3 Stations without any data in the "QUALITY" table

WaterbaseID	No. of stations
BA_RV_Kl1	1
BA_RV_U1	1
BA_RV_U2	1
BA_RV_U3	1
BA_RV_U4	1
BA_RV_Un1	1
ES_RV_ES01103	1
ES_RV_ES01109	1
ES_RV_ES01141	1
ES_RV_ES01353	1
ES_RV_ES05085	1
ES_RV_ES05087	1
ES_RV_ES05088	1
ES_RV_ES05130	1
ES_RV_ES08006	1
ES_RV_ES08015	1
ES_RV_ES08059	1
ES_RV_ES08089	1
ES_RV_ES08106	1
ES_RV_ES08144	1
ES_RV_ES08201	1
ES_RV_ES08207	1
ES_RV_ES08214	1
ES_RV_ES08216	1
ES_RV_ES08218	1

ES_RV_ES08231	1
FR_RV_03000450	1
FR_RV_03020145	1
FR_RV_03020941	1
FR_RV_03024460	1
FR_RV_03079570	1
FR_RV_03095200	1
FR_RV_04040200	1
GR_RV_GR_011220	1
GR_RV_GR_026620	1
GR_RV_GR_031020	1
GR_RV_GR_031030	1
GR_RV_GR_041310	1
GR_RV_GR_043020	1
GR_RV_GR_056020	1
GR_RV_GR_121130	1
GR_RV_GR_122030	1
SK_RV_P016000D	1
SK_RV_P112000D	1
SK_RV_S131010R	1
TOTAL	45

1.2.4 WaterbaseID in "PRESSURES" table not present in "STATIONS" table

WaterbaseID	No. of stations
TOTAL	0

1.2.5 Number of stations without any data in the "PRESSURES" table by country

Country code	No. of stations	Percentage of total no. of stations
BA	41	100
BE	3	7.14
BG	110	100
CS	77	100
DE	4	2.65
FI	192	97.46
FR	1	0.18
GB	14	6.86
GR	94	100
HR	45	100
HU	14	13.86
IS	1	100
IT	237	100
NL	11	47.83
RO	3	2.36
SK	5	8.33
TOTAL	852	23.73

2. Logical rule violation tests

Logical rules were tested in the “QUALITY” data table. This table contains several measurement value fields, calculated in 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
7	10Percentile >= Minimum
8	10Percentile <= Maximum
9	10Percentile <= 90Percentile
10	90Percentile >= Minimum
11	90Percentile <= Maximum
Rule	**Combined validation rules**
13	IF Minimum < Maximum THEN (StandardDeviation > 0)
14	IF NumberOfSamples = 1 THEN (Mean = Minimum = Maximum = Median = 10Percentile = 90Percentile)
15	IF NumberOfSamples = 1 THEN (StandardDeviation = 0)
16	IF NumberOfSamples = 0 THEN (AllValueType Is Null)
Rule	**Negative value validation rule**
17	All Values >= 0

Practice showed that some of rules needed small modifications which have been also applied:

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

Analysis of first results highlighted some obviously erroneous values in the data table, results of rounding errors mainly. Their removal significantly reduced total amount of the violations in the data table. Following cleaning formulas were used by EEA:

IF Minimum > 0 AND 10Percentile = 0 THEN change 10Percentile to Null
IF Minimum > 0 AND 90Percentile = 0 THEN change 90Percentile to Null
IF Minimum > 0 AND Median = 0 THEN change Median to Null
IF Rule 13 is violated THEN change StandardDeviation to Null

A special QA field (QA_LRviolations) has been added to the data table. Information of rules violated in respective record are holded there as a coma separated list of those rules numbers (numbers are same as in the table above). It is recommended the records where QA_LRviolation field is not empty (11602 records), should not be used in further analysis. Detected data quality inconsistencies will be tried to be solved in the near future.

3. Outlier detection

Detection of outliers were performed on the “QUALITY” data table. Following values were analyzed:

Measurement values: mean
Determinands: Nitrate, Total Oxidized Nitrogen, Orthophosphate, Total Ammonium, Total Phosphorus, BOD5, BOD7
Aggregation periods: annual

Measurement values were compared with other values from the same time series. If the value was detected as an outlier it was analyzed if it is an error or it was caused by natural conditions.

Some of the erroneous outliers have been corrected by EEA. Measurement values for some determinands might be provided by data suppliers in different units (e.g. *Nitrate* concentration might be provided in *mg N/l* or *mg NO_3/l*), but only one unit is used in published dataset. Values in other units have to be corrected by respective correction factors (e.g. *Nitrate* concentrations are published in *mg N/l*, values provided in *mg NO_3/l* unit have to be corrected by factor *0.226*). Some of the erroneous outliers were caused because the specific correction factor were not used or was used wrongly. Those values, where this error was detected, have been corrected and data suppliers will be informed:

Country: SK, Determinand: Nitrate, Year: 2004
- number of records with error: 43
- reason of error: correction factor 0.226 was not used because of the error in Unit field in the original delivery
- applied correction: all measurement values were corrected

Country: NL, Determinand: Nitrate, Year: 2005
- number of records with error: 22
- reason of error: correction factor 0.226 was wrongly used because of the error in Unit field in the original delivery
- applied correction: all measurement values were corrected

Records with remaining erroneous outliers (366 records), were marked in a special QA field (QA_outlier) which was added to data table. It is recommended the records where the value in QA_outlier field is TRUE, should not be used in further analysis. Also in these cases data suppliers will be informed.