Waterbase – Groundwater
Version 7

Quality control documentation

30 March 2007

Waterbase – Groundwater

Data on groundwater bodies 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.

Groundwater dataset include the physical characteristics of the groundwater bodies, proxy pressures on the groundwater area, as well as chemical quality data on nutrients and organic matter in groundwater.

QA/QC activities

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

Quality control tests have been performed on the Waterbase - Groundawater 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=987

Waterbase – Groundwater 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	1	100
BE	3	5.45
BG	5	5.95
CZ	4	9.3
ES	5	3.14
FR	902	68.85
GR	50	13.85
MK	7	100
PL	170	98.27
TOTAL	1147	46.82

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

WaterbaseID	No. of stations
TOTAL	0

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

Country code	No. of stations	Percentage of total no. of stations
AL	10	100
CS	11	100
CZ	43	100
DE	10	100
FR	35	2.67
GR	234	64.82
IS	1	100
IT	43	100
NL	9	100
PL	170	98.27
PT	7	70
TOTAL	573	23.39

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
7	10Percentile >= Minimum
8	10Percentile <= Maximum
9	10Percentile <= 90Percentile
10	90Percentile >= Minimum
11	90Percentile <= Maximum
Rule	**Negative value validation rule**
17	All Values >= 0

Analysis of first results highlighted some obviously erroneous values in the data table, results of rounding errors mainly. Their removal slightly 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 stored 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 (42 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

Initial detection of outliers was performed on the “QUALITY” data table. No obvious erroneous outliers have been detected but deeper analysis will be performed in the future. Therefore a special QA field (QA_outlier) has been added to the dataset. If erroneous outliers are detected, respective records will be marked there in future versions of the dataset.