European
Commission

Identification and treatment of outliers in the monitoring data for the monitoring-based exercise under the WFD Review of Priority Substances

Raquel N. Carvalho, Dimitar Marinov and Teresa Lettieri

Joint
Research
Centre

Table of contents

Background… 3

Possible interference of outliers with STE approach 3

Identification of outliers 3

Discussion 6

Conclusions. 6

References.. 7

Bentazone.. 8

This Technical Report is the JRC proposal to the SG-R group to remove the outliers of the monitoring data sets.

Raquel N. Carvalho, Dimitar Marinov and Teresa Lettieri

2015

Identification and treatment of outliers in the monitoring data for the monitoring-based exercise under the WFD Review of Priority Substances

(JRC proposal to SG-R)

1. Background

After collection of the monitoring data and performing summary statistics for each substance in the datasets, it is clear that for some substances, there are extremely high concentrations – for example in some cases in a range of g/L or several orders of magnitude higher than the limits of solubility of the substance. Possible explanations could be an erroneous reporting of the unit of measurement, typos, secondary contamination of samples, sampling or analytical problems, etc.

2. Possible interference of outliers with STE approach

The monitoring-based exercise of the present prioritisation process will be based on the STE approach, which calculates indexes for the **S**patial, **T**emporal and **E**xtent (**STE**) of PNEC exceedances in the monitoring data for each substance per observation station and sums them in a Final score as following:

$$\text{Final score} = F_{\text{spatial}} + F_{\text{temporal}} + F_{\text{extent}}$$

Initially, the STE method was developed intending to use maximum values per observation station for the calculation of both Fspatial and Fextent, assuming that datasets are "cleaned" of outliers (Ftemporal considers all records per site). However, as it became clear during the testing of the robustness of the STE method that this is not the casethe results will be distributed to the SG-R group in addition to this document), so, JRC is acting to update the STE methodology . The aim is to take into account the possible existence of outliers and try to decrease the influence of those outliers on the final score that will be used for the risk-based ranking of the substances.

It is acknowledged nevertheless, that for a few substances, the outliers may still affect the calculation of the STE factors. For this reason, the JRC is proposing an additional procedure for beforehand treatment of data intending to identify obvious outliers in the dataset and eventually remove them, in order to decrease the contribution of unrealistic concentration values to the final outcome of the monitoring-based prioritisation exercise. The rationale and method for the outlier identification and removal are described below.

3. Identification of outliers

Intuitively an outlier is an observation point that is distant from other observations. Outliers may be real values explained by the variability in the measured concentrations at different sites or days or may indicate experimental error or even a reporting error. There is however, no rigid mathematical definition of what constitutes an outlier and therefore, determining whether or not an observation is an outlier is ultimately a subjective choice. Nevertheless, there are various methods of outlier detection - some are graphical such as normal probability plots, others are model-based. Box plots are a hybrid of both.

The hybrid methods are usually built on the interquartile range. For example, if Q1 and Q3 are the lower and upper quartiles respectively, then one could define an outlier to be any observation outside the range [Q1 - k (Q3 - Q1), Q3 + k (Q3 - Q1)] where k is a positive constant.

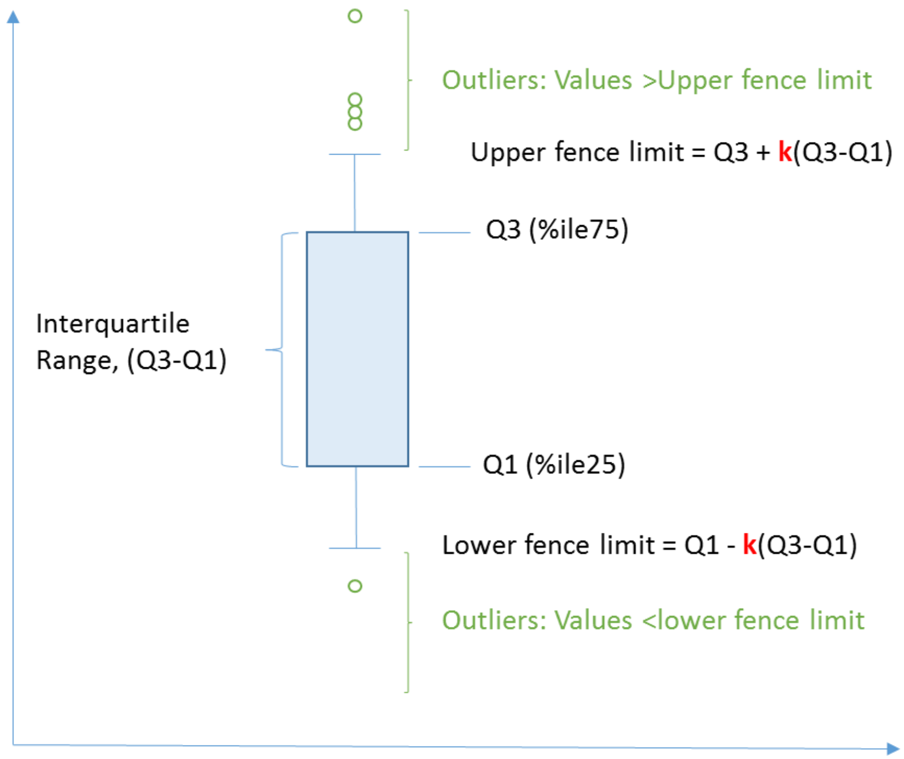

Outliers: Values >Upper fence limit
Upper fence limit = Q3 + k(Q3-Q1)
Q3 (%ile75)
Interquartile Range, (Q3-Q1)
Q1 (%ile25)
Lower fence limit = Q1 - k(Q3-Q1)
Outliers: Values <lower fence limit

Figure 1. Scheme for the distribution of measurements in a boxplot, and the identification of outliers by defining the fence limits with respect to the interquartile range.

For the identification of outliers under the monitoring-based prioritisation exercise, the JRC is proposing to use the method of interquartile range which is a box plot type as shown in the Fig.1 (see also the references). This is a simple and clear method that can be applied easily in R scripts. Moreover, the same method is used originally in R programming language (a software environment for statistical computing and graphics used by JRC) for making box-whisker plots (by choosing the constant k=1.5).

Three different scenarios for outlier identification have been tested by varying the value of the constant k - from the usual value 1.5 to 10 or 1000, thus changing the fence limits according to the equation displayed above. The upper and lower fence limits could then be used as threshold above and below which records may be deleted from the original dataset. It is interesting to note however, that in the monitoring dataset, the upper outliers are much more common than the lower outliers, and are considered to have a potentially higher impact in the final score of the substance. Thus, the upper outliers become the main target of the proposed procedure. Also JRC is open to discuss additionally alternative values of k-constant.

The proposed procedure for the elimination of outliers has been applied to the measurements of inland whole water dataset. Table 1 shows the counts of substances for which outliers are identified when the above mentioned three scenarios are considered, and the respective upper fence limits are compared with the maximum concentrations. For convenience the corresponding percentage from the total number of substances is also reported in the Table 1 (at this moment the total number of substances in the inland whole water dataset is 1692 but it is expected to be reduced as JRC is ready to apply an additional quality check of data). As can be seen the increase of k-value (from 1.5 to 1000) reduces the number of substances, identified to have upper outliers, more than 2 times (k=10 shows relatively low impact comparing to k=1.5).

A considerable decrease of the identified substances with upper outliers was observed if the outlier procedure uses per substance the 95-percentile of concentrations as an alternative to the maximum one (see Table1). Therefore, we found this option as too restrictive and JRC is recommending the maximum of concentrations to be used for comparison with the upper fence limit in the outlier procedure.

Table 1. Comparison of the maximum concentration or the 95-percentile of each substance with the upper fence value, which has been calculated using the constant k equal to 1.5, 10 or 1000, to identify the number of substances with outliers and the percentage with respect to the total number of substances. In red is the k value that JRC proposes to apply for the removal of outliers.

	k=1.5	**k=10**	**k=1000**
Max concentration > Upper fence limit			
# Substances	1196	990	522
% from the total number of substances	71	59	31
95th-percentile of concentration > Upper fence limit			
# Substances	735	303	163
% from the total number of substances	43	18	10

4. Discussion

The list of substances going through the monitoring-based exercise is very heterogeneous in terms of the chemical use, physical-chemical properties and fate in the environment. Therefore, a very broad range of concentrations may exist for different substances, which cautions against the adoption of a very small k value (k=1.5) for the identification of outliers.

Likewise, according to JRC the scenario k=10 is still too stringent since it would remove concentrations distant from the median that could be considered outliers for some substances, while for others, it is expected that there will be true heterogeneity in the measurements of the substance across Europe.

For this reason, the JRC proposes the use of k=1000 (this will eliminate the outliers for about 31% of substances). This assures the exclusion of most of the records for which a wrong unit of measurement has been reported, since this error often result in changes in the data of 10^3-fold or its multiples. While it is acknowledged that such an approach may not remove all outliers from the dataset, it does however remove the most extreme measurements that would have the highest impact in the STE approach, and at the same time the important variability information that is inherent to the heterogeneous chemical concentrations across Europe is retained.

As an illustration, in the figure below it is shown the application of the procedure for Bentazone for which the outliers issue was identified and discussed during the call conference for updating its dossier. The outliers that would be removed from the dataset are labelled red. In the figure the Scenario 2 is related to the dataset (not to the outlier procedure) and indicates all measurement data, while Scenario 1 corresponds only to the quantified records (value > LoQ or LOD). Bentazone has been chosen just as example, we identified as well outliers for other substances for which the dossiers have been updated. Moreover, it is important to remark that the outlier removal approach will be applied also to

1. The substances, highly ranked in the last prioritisation exercise, which dossiers have been updated by integrating as well the monitoring data to compare the provisional PNEC with the measured concentrations to evaluate the exposure risk (exceedance of the provisional PNEC).
2. The monitoring based exercise based on STE approach in the ongoing prioritisation process.

Conclusions

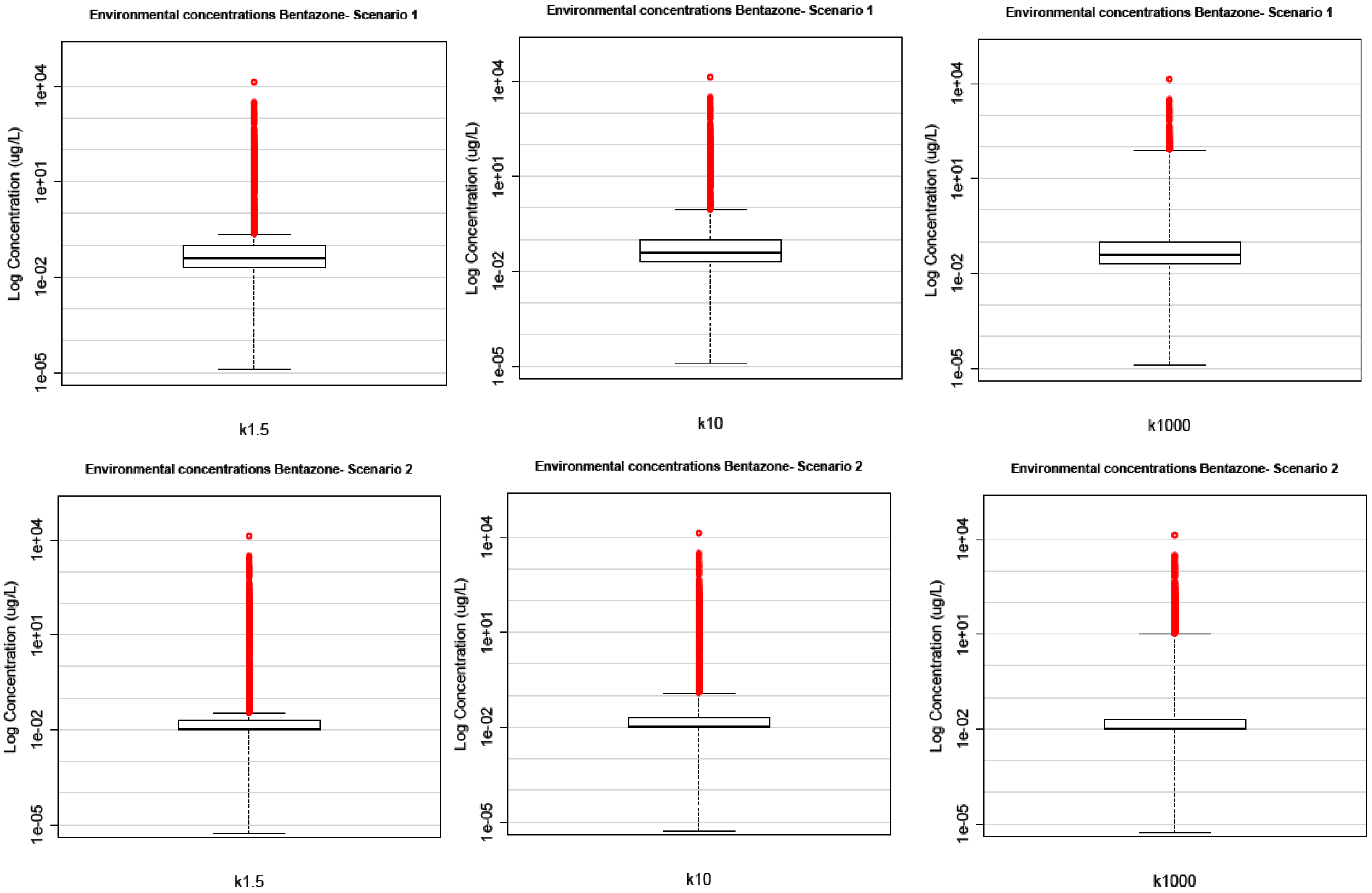
To reduce the influence of outliers of the monitoring data sets, JRC proposes two steps , one shortly introduced in section 2 (update the STE approach which will be extensively explained in a separated document to be distributed soon) and the other described in section 3 (identification and removal of outliers) and discussed in the section 4.

The combination of these two actions ((outliers removal following the approach described above with a k=1000), together with the updated STE approach for the monitoring based exercise, will assure that the highest-ranks will be well correlated with the highest risk imposed by the substances in Europe.

References

1. Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.
2. Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) Graphical Methods for Data Analysis. Wadsworth & Brooks/Cole.
3. Murrell, P. (2005) R Graphics. Chapman & Hall/CRC Press.

Bentazone: identification of outlier as values falling outside the range [Q1 - k (Q3 - Q1), Q3 + k (Q3 - Q1)]. Three variants have been tested considering different k values (1.5, 10 or 1000). The outliers that would be removed from the dataset in each case are labelled red. The Scenario 2 is related to the dataset (not to the outlier procedure) and indicates all measurement data, while Scenario 1 corresponds to quantified values only (value > LoQ or LOD).


Environmental concentrations Bentazone- Scenario 1
Log Concentration (ug/L)
1e+04
1e+01
1e-02
1e-05
k1.5
Environmental concentrations Bentazone- Scenario 1
Log Concentration (ug/L)
1e+04
1e+01
1e-02
1e-05
k10
Environmental concentrations Bentazone- Scenario 1
Log Concentration (ug/L)
1e+04
1e+01
1e-02
1e-05
k1000
Environmental concentrations Bentazone- Scenario 2
Log Concentration (ug/L)
1e+04
1e+01
1e-02
1e-05
k1.5
Environmental concentrations Bentazone- Scenario 2
Log Concentration (ug/L)
1e+04
1e+01
1e-02
1e-05
k10
Environmental concentrations Bentazone- Scenario 2
Log Concentration (ug/L)
1e+04
1e+01
1e-02
1e-05
k1000