COMPARING TWO METHODS FOR QUALITY CONTROL OF POPS DATA IN INTERCALIBRATION STUDIES

Libralato S[1] , Raccanelli S[2], Van Bavel B[3]

[1] Istituto Nazionale di Oceanografia e di Geofisica Sperimentale - OGS, Borgo Grotta Gigante 42/c, 34010, Sgonico (TS), Italy; [2] Consorzio I.N.C.A., VEGA-Edificio Cygnus, Via delle Industrie 21/8, 30175 Marghera (VE), Italy; [3] MTM Research Center, School of Science and Technology, Örebro University - SE 701 82 Örebro, Sweden.

Introduction

The multiannual experience of Circuit Inter-laboratory for Dioxins (CIND)[1] and INTERCAL[2], highlighted the importance on objective methods for quality control and quality assurance (QC/QA) of laboratories data in intercalibration and intercomparison studies. In particular, large errors caused by misreporting (including unit of measure) or rough laboratory mistakes might sum to the noise to produce skewed and dispersed values that are difficult to detect objectively and automatically. Expert-based evaluation might help considerably in detecting these large errors but this data-check is necessarily subjective[3]. Therefore the setting of an objective and standard method, other than speeding-up the QC/QA process, might overcome all the problems due to the subjectivity of the large errors identification and might permit a serious standardized identification of outliers and extremes.
Here we compare two methods for QC/QA that have been used in intercalibration studies and in other similar cases [1,3]. Both methods identify the outliers and extremes on the basis of statistical indices calculated on the provided POPs data. The two methods differ because one method is based on parametric statistics, i.e. mean and standard deviation (called in the following P method) and the other is based on non-parametric statistics such as median and quartiles (called in the following NP method). The two methods are compared by using synthetic dataset with a known true value, random noise of fixed dispersion and random large errors whose proportion is controlled. The comparison will permit to identify pros and cons of the two methods in identify the true value of the synthetic POPs dataset.

Materials and methods

Synthetic dataset.

Several dataset were generated to resemble real conditions. Therefore dataset were generated by assuming that a number of laboratories (N= 50, 100, 150) produced POPs concentrations for PCDD/F (17 congeners), PCB (12 congeners) and PAH (7 congeners). True values for POPs concentration were set at values ranging from 0.0001 ng/g to 1000 ng/g, thus covering 7 order of magnitude of concentration. Random noise in the data was added by generating, for each of the N laboratories, random values normally distributed with known dispersion (relative standard deviation, RSD%=20% as reported elsewhere [2]) around the true value. The numbers were generated with an opportunely created VisualBasic macro that implements the RAN3 algorithm[4] that allows for avoid problems of usual number generator. Moreover, at random were added rough large errors of two types: a) one interesting all the congeners measured by the randomly chosen laboratory; b) one error randomly placed in the matrix of data. Error a) represent the possibility that the lab misreport the data entirely for different reasons (problem in the conservation of the sample, error in the evaluation of the weight or humidity of the sample; error in unit of measure); error b) represent a true rough mistake, possibly occurring for the same reasons but only for that congener. Number of lab with a) error were defined a priory but chosen at random, and assuming the true values is missed by a factor of 10. Number of data with b) error were defined a priori but chosen at random and assuming true value was missed by a factor of 1000. Several datasets were created to represent realistic conditions in intercalibration studies by varying the number of labs (N= 50, 100, 150), the true value for the congeners (from 10^{-4} up to 10^{3}), the RSD% from 10% to 30%, the proportion of a) and b) large errors. Here results are synthetically reported for some of the experiments.

Quality control methods

The values of the dataset $x_{i,j}$, were i represent the laboratory and j the congener, were used to calculate statistical indices. For each congener (j) were calculated across the synthetic dataset: average concentration ($\overline{x}_j$),

standard deviation (s_j), coefficient of variance (RSD%j=$s_j / \bar{x}_j$ %), the median (Mj) and the 1st and 3rd quartiles (Q25j and Q75j respectively).

In the P-Method, outliers and extremes were identified according with the following criterion:

$$\bar{x}_j - 2 \cdot s_j > x_{i,j} \quad \text{or} \quad x_{i,j} > \bar{x}_j + 2 \cdot s_j \qquad \text{(Eq. 1)}$$

The identified outliers were removed for the evaluation of the true value and the calculation of other indicators (e.g. z-scores). However, given the large distribution of data, in some cases were necessary to reiterate the application of the method. Therefore after the first run for removing outliers, statistical indexes were recalculated and new outliers identified and removed. Therefore the procedure for removing outliers was repeated twice in cases were necessary[1].

Conversely in the NP-method the non-parametric indexes (median and quartiles) were used to indentify extremes and outliers. In particular, the range between the two quartiles for each congener was defined as:

$$U_j = Q75_j - Q25_j \qquad \text{(Eq. 2)}$$

and extremes and outliers were identified as the values that are outside the range defined as[2]:

extremes: $Q25_j - 3 \cdot U_j > x_{i,k,j}$ and $Q75_j + 3 \cdot U_j < x_{i,k,j}$ (Eq. 3)

outliers: $Q25_j - 1.5 \cdot U_j > x_{i,k,j}$ and $Q75_j + 1.5 \cdot U_j < x_{i,k,j}$ (Eq. 4)

Extremes and outliers identified were excluded from the data set for calculations of the "true value" and statistical indexes calculated across laboratories. In figure 1 it is schematically represented the definition of outliers and extremes in the two methods. The analysis revealed the potential differences, capabilities and limitations of the two methods. Performances were evaluated by comparing the sum of squares of differences between log concentration (SSlog) of treated and true values for all congeners (17 PCDD/F, 12 PCB, 7 PAH). The SSlog was also calculated for input data (synthetic dataset) and true values (average concentration) to show the original dispersion of data and the improvement. Moreover, methods were also compared in terms of improvement of RSD% from synthetic dataset to treated dataset after the two methods were applied.

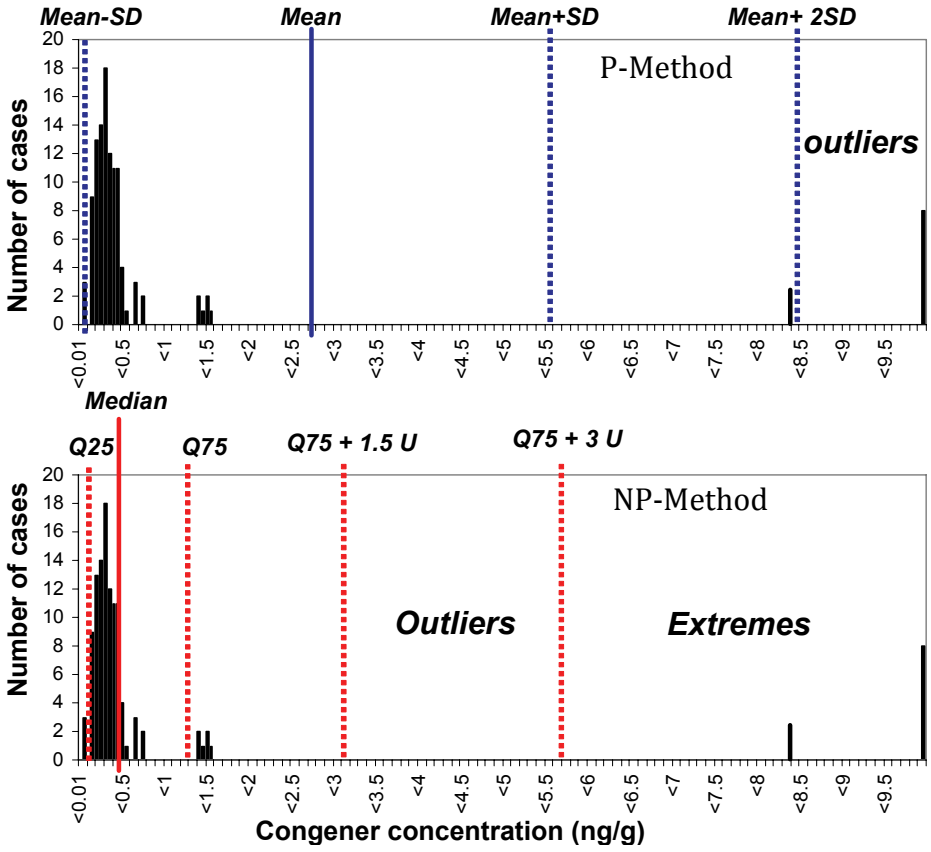

Figure 1. Exemplification of the two methods used for QC/QA in a synthetic congener data distribution: the parametric (P-mehtod, above) and the non-parametric (NP-method, below). Identification of the ouliers and extremes is reported.

Results and Discussion

The input data (synthetic dataset) had dispersion that resemble that of a interlaboratory circuit, with data often not normally distributed and extreme dispersed values (RSD% up to 600%). The SSlog between true value and synthetic dataset TEST1 (150 laboratories, RSD% 20%, including 5 randomly chosen lab with all data wrong by a factor of 10 and 40 data with large errors of a factor of 1000), was 4.204. The application of the P-Method implied the reduction of SSlog to 0.403 (intermediate data in Figure 2), but dispersion was still high. It was thus necessary to repeat the P-method twice to eliminate all the large errors data and reach a SSlog = 0.014. This goal was met in one application of NP-method over the TEST1 dataset, as can be seen in Figure 2 (lower panels): starting from the same dataset the NP-Method identified outliers and extremes very efficiently and reduced SSlog to 0.014.

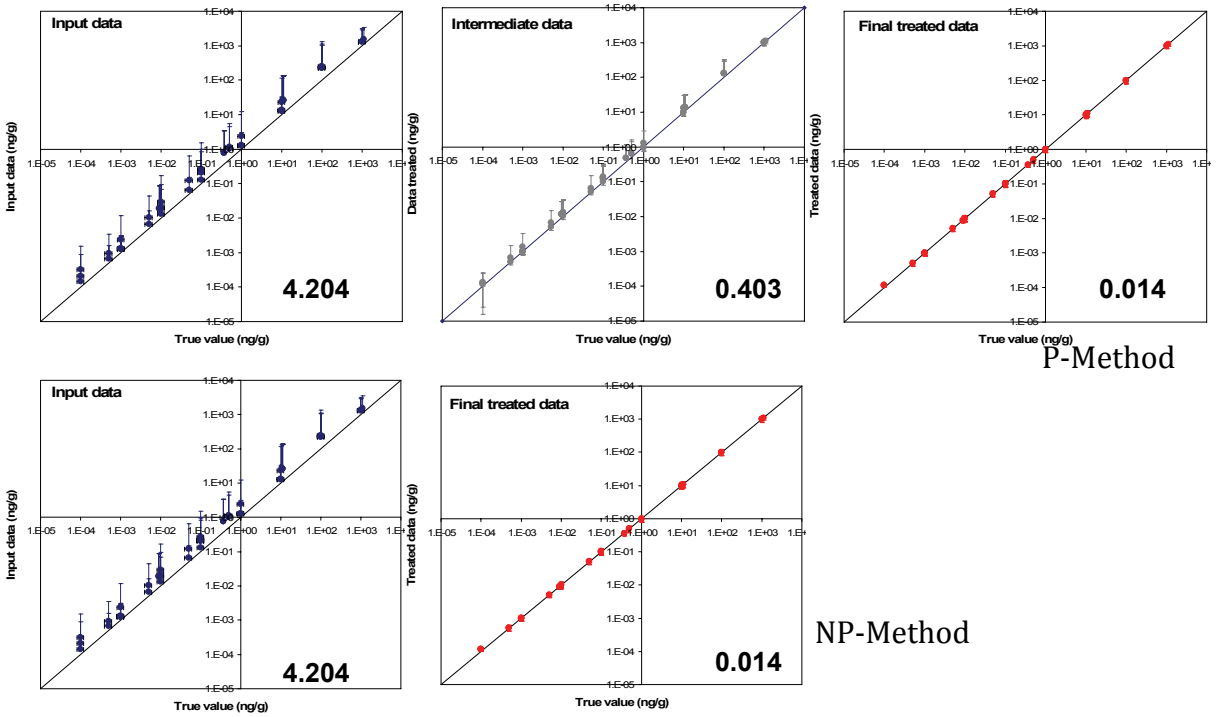

Figure 2. Schematic representation of data (mean and standard deviation in log scale) of input dataset synthetically generated (left) for dataset TEST1 (see text). Results of application of P-Method (upper panels) and NP method (lower panels). The value in the low right part of each panel represents the SSlog value (sum of squares of differences between log concentration).

Figure 3 allows to highlight that the better efficiency of NP-Method is also with regards to RSD%. As can be seen from Figure 3, in fact, P-Method applied to TEST1 dataset allowed to reduce the RSD% to the noise value (20%), but for P-method it was necessary to apply it twice, reiteratively. Table 1 synthesized the tests conducted using different synthetic datasets, with different number of synthetic laboratories, with different proportion of large errors (type a and type b). The analysis conducted revealed the better performances of NP-Method, whereas P-method has to be applied twice to reach the same performances in terms of closeness to true value and dispersion. However, should be noted that N-method, might also fail in cases of absence of large errors, by identifying noise error data as outliers (TEST 7) and thus reducing below true value the dispersion of treated data.

This preliminary analysis reveals the importance of using non parametric analyses in QA/QC since parametric measures, by assuming symmetric distribution might fail to represent data and to detect outliers (see for instance Figure 1). This analysis might represent a first step toward the setting of a common methodology to be used in QA/QC of intercalibration studies [1,2].

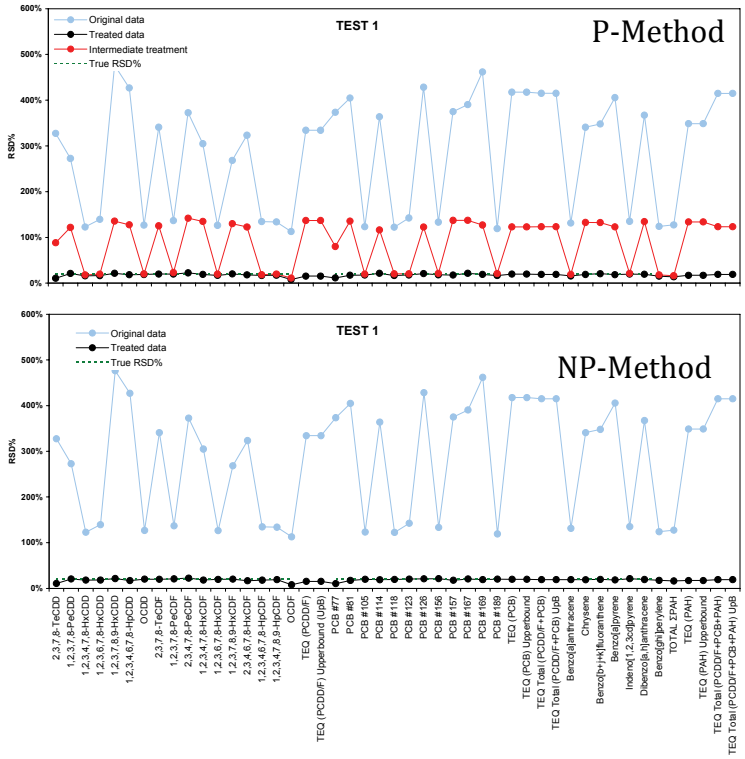

Figure 3.
Schematic representation of coefficient of variation for all congeners and derived concentrations (such as TEQ) over laboratories for TEST1 input dataset and treated data for P-Method applied twice (above) and NP-Method (below). The figures provide basis for evidencing the accuracy related to the evaluation of the true value for each congener.

Table 1.
Results of the application of the two methods for QC/QA on different synthetic datasets, created with different number of lab and different proportion of large errors of type a) and b). Performances are evaluated in terms of SSlog with known true value.

Dataset	N lab	Values range	RSD%	N lab Err (a)	N data Err (b)	Method	SSlog Input	SSlog QC1	SSlog QC2
TEST 1	150	10^{-4} - 10^{3}	20%	5	40	P	4.204	0.403	0.014
						NP	4.204	0.014	
TEST 2	100	10^{-4} - 10^{3}	20%	3	26	P	6.599	0.565	0.016
						NP	6.599	0.015	
TEST 3	50	10^{-4} - 10^{3}	20%	2	13	P	1.691	0.023	0.020
						NP	1.691	0.020	
TEST 4	150	10^{-4} - 10^{3}	20%	0	140	P	8.565	0.019	0.012
						NP	8.565	0.014	
TEST 5	150	10^{-4} - 10^{3}	20%	6	57	P	41.301	0.561	0.013
						NP	41.301	0.014	
TEST6	150	10^{-4} - 10^{3}	20%	0	63	P	39.196	0.014	0.012
						NP	39.196	0.013	
TEST 7	150	10^{-4} - 10^{3}	20%	0	0	P	0.015	0.013	0.012
						NP	0.015	0.014	
TEST 8	150	10^{-4} - 10^{3}	20%	5	0	P	0.676	0.015	0.013
						NP	0.676	0.014	

References:
1. Raccanelli S., Petrizzo A., Favotto M. and Pastres R. (2007). *Organoh. Comp*. 69: 982-985.
2. van Bavel B., Abad E., 2008. *Anal. Chem*. 80 (2008) 3956-3964.
3. Raccanelli S., Libralato S., 2009. *Organoh. Comp*. 71: pp 336-340.
4. Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T., 1987. *Numerical Recipes, the art of scientific computing*. Cambridge University Press.