Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 533 (2005) 31–39
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

A new model for the inference of population characteristics from experimental data using uncertainties Part II. Application to censored datasets

Wim P. Cofino[a,*], Ivo H.M. van Stokkum[b], Jaap van Steenwijk[c], David E. Wells[d]

[a] *Wageningen University, Environmental Sciences Group, Subdepartment of Water Resources, Hydrology and Quantitative Water Management Group, Nieuwe Kanaal 11, 6709 PA Wageningen, The Netherlands*
[b] *Department of Physics Applied Computer Science, Division of Physics and Astronomy, Faculty of Sciences, Vrije Universiteit, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands*
[c] *Ministry of Transport, Public Works and Water Management, Directorate General for Public Works and Water Management, Institute for Integrated Inland Water Management and Waste Water Treatment RIZA, P.O. Box 17, 8200 AA Lelystad, The Netherlands*
[d] *Fisheries Research Services, Victoria Road, Aberdeen, UK*

Received 4 April 2004; received in revised form 2 November 2004; accepted 2 November 2004
Available online 15 December 2004

Abstract

This paper extends a recent report on a model to establish population characteristics to include censored data. The theoretical background is given. The application given in this paper is limited to left-censored data, i.e. *less than* values, but the principles can also be adopted for other types of censored data. The model gives robust estimates of population characteristics for datasets with complicated underlying distributions including *less than* values of different magnitude and *less than* values exceeding the values of numerical data. The extended model is illustrated with simulated datasets, data from interlaboratory studies and temporal trend data on dissolved cadmium in the Rhine river. The calculations confirm that inclusion of left-censored values in the computation of population characteristics improves assessment procedures.

Keywords: Censored data; Censored samples; Left-censored data; Maximum likelihood estimation (MLE); Detection limit; Robust statistics; LOD; LOQ; *Less than* values; Mean; Standard deviation; Interlaboratory studies; Proficiency testing

1. Introduction

Censored data, i.e. datasets that include non-numerical values, are frequently encountered in different fields of science [1–4]. The non-numerical values may be known to be below a certain limit, e.g. left-censored data as *less than* values, and/or above an upper limit. Values below a limit of quantification (LOQ[1]) are frequently encountered both in environmental studies and in interlaboratory studies [5–7]. Assumptions need to be made if these *less than* values are to be incorporated into the calculation of the population characteristics. Apart from the removal of *less than* values from the dataset, a common approach is to substitute the *less than* values by a constant value like the LOQ itself, half the LOQ or zero. The most widely accepted and recommended substitution is half the LOQ. However, several studies have shown that simple substitution methods perform poorly in comparison to other methods in summary statistics [8–10]. In order to improve the estimate of the summary statistics, methods have been developed that combine the numerical values with extrapolation of below-limit values, assuming a specific probability density function (pdf). The maximum

* Corresponding author. Tel.: +31 317 474304; fax: +31 317 484885.
E-mail address: wim.cofino@wur.nl (W.P. Cofino).

[1] This paper uses LOQ to denote the limit that is reported when *less than* data are encountered. Values above this limit are referred to as 'numerical data'.

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.11.008

likelihood method and log probability plotting are two examples [11]. In many environmental datasets *less than* values occur along with potential outliers in the right hand tail of the distribution. Robust estimation techniques have been developed to deal with such situations [3].

Recently, a new model to calculate the population characteristics for experimental data has been reported [12]. This model does not assume unimodality of the distribution and provides a robust estimation of population characteristics. In this paper, the development of this model is described which includes left-censored values. Following an outline of the theory, the approach is illustrated with calculations on simulated datasets, on data from interlaboratory studies and on data from water quality monitoring. The model can be adapted in the same manner to include other types of censored data.

2. Theoretical background

Data arise from a measurement process which, when under control, gives an output that can be described by a specific probability density function (pdf). A pdf can be attributed to a particular dataset by adding up the pdf's associated with all the individual independent measurements. The overall pdf constructed in this manner is the starting point for the model. Instead of calculating the mean of the data, the model sets out to establish the most probable value given the overall pdf. The mathematical procedure borrows the concept of wavefunctions from quantum mechanics. This enables the use of powerful matrix algebra. As an analogue to wavefunctions, observation measurement functions (OMF, φ_i)[2] are defined as the square root of the probability density function which is attributed to the individual observation in question. The set of OMFs forms a space, or a basisset, in which so called population measurement functions (PMFs[2]) are constructed. The construction of the PMF Ψ_i is a linear combination of OMF's, i.e. $\Psi_i = \sum c_{ij}\varphi_j$. A normalised, squared PMF is a pdf.

In the model, the coefficients c_{ij} are obtained by seeking for the (unnormalised) PMF which has the highest probability in the basisset. The probability of PMF_i is obtained as the integral $\int \Psi_i^2 \,\mathrm{d}x$. Mathematically we have to establish the set of coefficients for which the integral $\int \Psi^2 \,\mathrm{d}x$ is maximal. The mathematical procedure uses the method of Lagrange multipliers and imposes the additional constraint, that the sum of the squared coefficients is equal to one.

The mathematical elaboration requires a solution to the eigenvector–eigenvalue equation $Sc = \lambda c$. In this equation, S represents the matrix of overlap integrals. For example, the matrix element S_{12} is calculated as $\int \varphi_1\varphi_2 \,\mathrm{d}x$, i.e. the integral of the product of OMF_1 and OMF_2. S_{12} provides a quantitative measure how well the two observations agree, taking the respective pdf's into account. The overlap integral can range between 0 (no overlap) and 1 (100% overlap) when the observations have identical pdf's.

The model renders in a set of n basisvectors OMF a total of n eigenvectors with eigenvalues λ. The eigenvalue λ_i gives the probability in the basisset of the corresponding eigenfunction i. The highest probability and thus maximum value for λ is equal to the number of data n, which is obtained when all data have exactly the same pdf. In this case, each OMF has a coefficient which is equal to $1/\sqrt{n}$. The eigenvector with the highest eigenvalue λ is the PMF_1. The remaining $n-1$ linear combinations are ranked according to probability (i.e. eigenvalue) and are denoted as $\mathrm{PMF}_2, \ldots, \mathrm{PMF}_n$. PMF_2 and higher PMF's may sometimes be additional modes, but are frequently only clusters of data ordered according to their degree of overlap. Each squared PMF effectively describes a part of the pdf of the ensemble of data. When the squared PMFs are summed together over the entire concentration range, the pdf of the entire dataset is reconstructed.

For each PMF Ψ_i the expectation value and variance can be calculated as follows:

$$m_i = \frac{\int x\Psi_i^2 \,\mathrm{d}x}{\int \Psi_i^2 \,\mathrm{d}x}, \qquad s_i^2 = \frac{\int x^2\Psi_i^2 \,\mathrm{d}x}{\int \Psi_i^2 \,\mathrm{d}x} - \bar{m}_i^2$$

In addition to the mean and standard deviations of each mode or cluster, the eigenvalues λ enable the quantitative assessment of the degree of comparability and the character (unimodal, bimodal) of the dataset. To this end, the program converts the eigenvalue of the mode or cluster proper into a percentage of the overall pdf. The percentage therefore quantitatively describes which fraction of the dataset is accounted for by the PMF in question.

The model is extended for use with *less than* values by applying the appropriate probability density functions. A straightforward approach can be taken when no assumptions are made regarding the probability density function underlying a *less than* value. In such a case, in a first approximation each concentration between zero and the LOQ has an equal probability. We can then use the square root of a rectangular probability density function as basisfunction. Explicitly, when a *less than* value is reported, the basisfunction is equal to $\sqrt{1/\mathrm{LOQ}}$ in the interval between zero and LOQ and zero otherwise. These basisfunctions have an expectation value $m_i = \int \varphi_i^2 x \,\mathrm{d}x = \mathrm{LOQ}/2$ and a variance $\int \varphi_i^2 x^2 \,\mathrm{d}x - \bar{m}_i^2 = \mathrm{LOQ}^2/12$. When specific knowledge of the measurement process and the properties of the measured object is available, it would be possible to use other probability density functions. Montville and Voigtman derived pdf's for the instrumental limit of detections [13]. These pdf's can be used when the model is specifically applied to such data. The implicit assumption made with the maximum likelihood method and log probability plotting techniques entails that the LOQs are cut off from the population formed by the numeri-

[2] In this and following papers, the terminology is changed somewhat in comparison with reference [12]. Laboratory measurement function is replaced by observation measurement function, interlaboratory measurement function is now denoted as population measurement function. This modification is applied as the scope of the model is much broader than interlaboratory studies.

cal values, implying that a concentration just below the LOQ is more likely rather than near zero. To mimic this assumption in a simple way, in this paper a basisfunction has been defined as the square root of a simple triangular pdf. This triangular pdf has the form $(2/\mathrm{LOQ}^2)x$ for concentrations between zero and LOQ and zero otherwise, with an expectation value $m_i = \int \varphi_i^2 x\,\mathrm{d}x = 2 \times (\mathrm{LOQ}/3)$ and a variance $\int \varphi_i^2 x^2\,\mathrm{d}x - \bar{m}_i^2 = \mathrm{LOQ}^2/18$.

Recently, the kernel density approach has been proposed to study the features of the population [14]. In this method, each datapoint is assigned a normal distribution with a fixed standard deviation. This standard deviation is obtained using the h-estimator, which is optimised so as to obtain a meaningful appearance of the graphical representations of the population.

As with the kernel density approach, our model uses pdfs as building blocks. The key difference lies in linking the pdfs to the concept of measurement functions and by using matrix algebra to calculate the features of the population as outlined above. The model has an implementation, the normal distribution approximation (NDA), which does not require the individual uncertainties of the datapoints [12]. In this implementation each observation is attributed a normal distribution with one and the same standard deviation. This standard deviation is estimated so as to reproduce the population characteristics of a normal distribution quantitatively. The kernel density method and the normal distribution approximation of the model produce very similar graphs of the population. The kernel density approach and our model are complementary, however our model provides additional tools for exploratory data analysis (e.g. graphical representation of the overlap matrix, see Fig. 2 of the paper, and plots of the eigenvectors, see [12]) as well as the quantitative results in addition.

The model is very flexible and can be applied in various ways both with respect to the type of probability density functions, e.g. normal distributions, Students *t*-distribution, rectangular distributions, and the uncertainty characteristics, e.g. standard deviations reported by laboratories or a common standard deviation.

The program [12] has been extended to include *less than* values. Integrals between basisfunctions invoking the product of the square root of a normal distribution respectively a rectangular or triangular pdf as described above are obtained by numerical integration. Integrals among the rectangular or the triangular functions are carried out using the analytical functions. Integrals among basisfunctions based on the normal pdfs are obtained as previously reported. The program is provided as a free Matlab toolbox upon request.

3. Comparison of methods on simulated datasets

The extended model is demonstrated using a simulated dataset following the approach described by Kuttatharmmakul et al. [2]. A total of 250 datasets consisting of twelve observations were generated from a normal distribution with mean 1.09 and standard deviation 0.20. Subsequently, observations less than one were treated as a *less than* value with LOQ = 1. Only datasets with at least one LOQ were included in the calculations. The means and standard deviations for each dataset were calculated with two methods: the Cohen maximum likelihood method estimator [2] and the model using a rectangular pdf for the LOQs. The Cohen maximum likelihood method was selected since it is regarded as an appropriate approach to incorporate left-censored data into the evaluation [2]. The main restriction in the use of the method is that it requires the data to be normally distributed and it can only accept one value for the left-censored data. The results of the calculations are depicted in Fig. 1.

The Cohen maximum likelihood estimator and the model give comparable results when the number of *less than* values is below five. The two methods disagree when the number of *less than* values exceeds five. The Cohen maximum likelihood estimator requires the numerical data also at high LOQ percentages to estimate the characteristics of the assumed underlying distribution and thus to calculate mean and standard deviation adjusted for LOQs. The model does not invoke any assumption about the character of the overall population. When more than five LOQs are present, the model indicates that the dataset is bimodal. The first mode consists of the six or more *less than* values which all have the same pdf. In principle, the expectation value of this mode is 0.5 (i.e., the expectation value of the individual basisfunctions). Higher expectation values occur when numerical data with a value close to one are present. Such data have pdfs that overlap with the pdfs of the *less than* values. Because of this overlap, the expectation value of the first mode is increased. The second mode consists of the numerical values. In a conventional interpretation, the model indicates that the numerical data are outliers when the number of *less than* values is greater than five. For an interlaboratory study, the interpretation might be that the higher values are attributed to false positives.

When the number of *less than* values equals five, the level of agreement between the Cohen maximum likelihood esti-

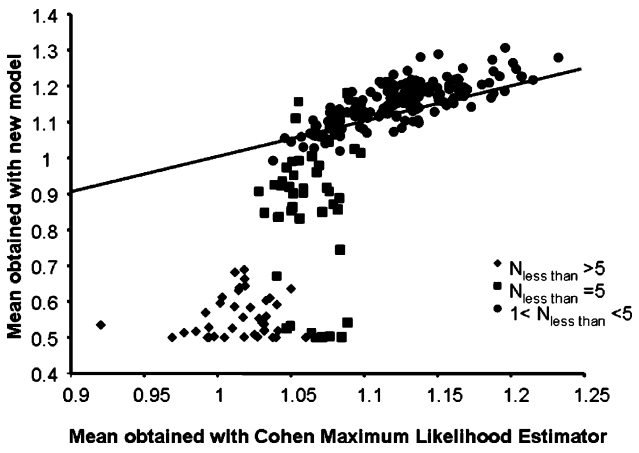

Fig. 1. Results of two methods to calculate the mean of left-censored data on 250 simulated datasets. The line $y = x$ is drawn to facilitate the comparison.

Table 1
Calculations on polybrominated flame retardants (data from De Boer and Cofino, 2002)

Matrix and congener	All data				Only numerical data				Numerical data and LOQs ≤ NDA mean of numerical data[a]			
	Nobs	Expectation value	S.D.	%	Nobs	Expectation value	S.D.	%	Nobs	Expectation value	S.d.	%
Eel—BDE 209	12	0.78	2.49	35.2	4	0.078	0.083	34.6	5	0.074	0.082	27.9
Eel—BDE 119	9	0.042	0.048	51.9	4	0.038	0.023	52.6	4	0.038	0.023	52.6
Mussel—BDE 153	11	0.034	0.034	39.1	7	0.047	0.018	44.4	9	0.037	0.019	37.1
Cormorant—BDE 66	8	0.15	0.14	44.2	3	0.063	0.018	50.1	5	0.039	0.024	37.5
Porpoise Liver—BDE 209	13	4.71	8.12	37.1	4	7.50	3.17	47.2	10	1.59	1.77	36.6
Sediment_7—BDE 75	6	0.036	0.045	40.0	4	0.26	0.132	44.5	6	0.036	0.045	40.0

[a] The NDA mean of the numerical data is the expectation value of PMF_1 obtained by applying the normal distribution approximation (NDA) implementation of the model to the numerical data. The NDA approach does not require the specification of the uncertainties of the laboratories [12].

mator and the model varies significantly. This can be traced back to the characteristics of the dataset. Depending on the distribution of the numerical data the first mode is made up by the numerical data, the *less than* data, or by a combination of both. In the first case, a good correspondence with the Cohen method is obtained. In the latter two cases, the agreement with the Cohen method is less good.

The calculations indicate that the Cohen maximum likelihood estimator and the model give comparable results except when the number of *less than* values is high. This difference arises as the approaches are based on different principles. The Cohen method assumes a normal distribution for the numerical data and corrects for the *less than* values. Our model sets out to calculate the performance characteristics of the 'first mode' of the dataset, regardless whether this mode is composed of numerical or censored data. The availability of statistical methods based upon different principles is an advantage. When the outcomes of the methods differ, the dataset should be inspected. It should be judged whether the assumptions underlying the statistical methods are met. The nature of the measurement should be taken into account—are measurement problems (e.g. contamination, incomplete resolution) possible? The statistical procedures have thus to be complemented by chemical expert judgement. This judgement will determine whether it is possible to make a statement about the performance characteristics of the dataset at all.

4. Case study I—interlaboratory study on polybrominated diphenylethers (PBDEs)

The result of a recent interlaboratory study on PBDEs has been reported [5]. Datasets in this study contained a small number of observations with a relatively high number of left-censored data which varied considerably in magnitude. The numerical data exhibit a wide scatter and had difficult underlying distribution profiles. A selection of the data from this study are used to illustrate the extended model.

Initially, calculations were made with the full dataset and then with the dataset without the *less than* values. In most cases, inclusion of the *less than* values had a small effect. In Table 1, results are given for some difficult datasets. For BDE 119 and 209 in eel and BDE 66 in cormorant the calculations on the full datasets, including all *less than* values, give a higher expectation value than the calculations on the datasets from which all the *less than* values have been removed. This pattern is caused by *less than* values with high LOQs. This effect is illustrated for BDE 209 in eel with a graphical representation of the overlap matrix given in Fig. 2. The numerical data exhibit a poor comparability (observations 9–12 in Fig. 2). In this case, the model gives an expectation value of 0.078 ± 0.082 for the first mode, representing 34.6% of the dataset. This expectation value is determined predominantly

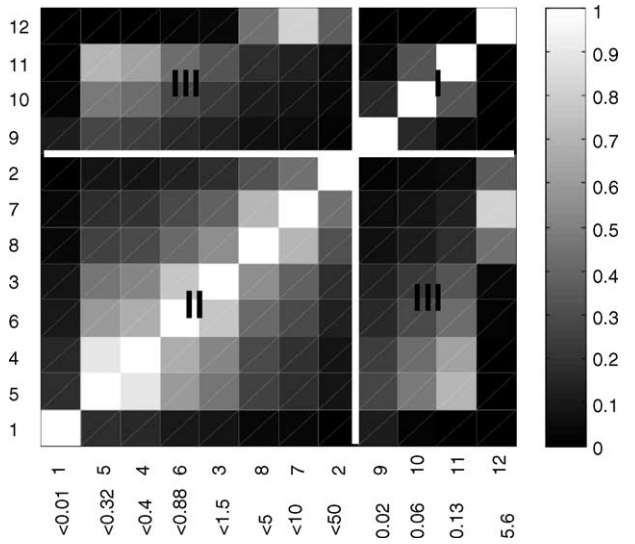

Fig. 2. Graphical representation of the overlap matrix for BDE 209 in eel. The overlap integral can have values between 0 and 1. The bar on the right depicts the relationship between gray scale and the magnitude of the overlap integral—white represents an overlap of 1, black represents no overlap. The observations 9–12 are numerical data, the observations 1–8 are *less than* values. The observations are ordered according to their magnitude. The figure has been divided into three zones defined by Roman numerals I, II and III. I is a 4×4 matrix of the numerical data, II is a 8×8 matrix depicting the overlaps between the *less than* values and III depicts the overlaps between the numerical data and *less than* values. The row of data at the bottom of the figure provides the concentrations reported.

by the observation 10, which overlaps moderately with both the observations 9 and 11 (overlaps respectively 0.16 and 0.34). The expectation value for the entire dataset including all *less than* values is calculated to be 0.78 ± 2.5, which accounts for 35.2% of the dataset. The 10-fold increase in expectation value is due to the rise of a new cluster with strongly overlapping data along with the introduction of the *less than* values. This cluster includes the LOQs < .32, < .4, < .88 and < 1.5 (observations 5, 4, 6, and 3 in Fig. 2). Similar effects occurs with the introduction of LOQs into the calculations for BDE 119 in eel and BDE 66 in cormorant. This observation suggests that the magnitude or the indicative information of a LOQ is important in any assessment. Clearly, the indicative information of LOQs which are an order of magnitude or more greater than numerical data is virtually zero. An example is the LOQ of <50 for BDE 209 in eel is which substantially greater than the expectation value of 0.078 based on the reported numerical data. The degree to which the calculations are affected by the high LOQs depends on the nature of the dataset. When there is a large number of laboratories reporting numerical data that are in good agreement amongst themselves, the presence of a limited number of high LOQs only has a small effect. Effects become greater when there is a small number of numerical data and several 'high' LOQs occur which overlap with themselves and/or with numerical outliers. LOQs higher than the median of the numerical data probably contain the true concentration, but provide little information and may perturb the calculations.

In this paper the calculations have been repeated with a constraint on the magnitude of LOQs which can be accepted. The constraint imposed was that only LOQs are included which are equal to or less than the expectation value obtained for the set of numerical data with the normal distribution approximation of the model [12]. The advantage of this approach is that an unwanted effect on the calculations arising from high LOQs is prevented. The disadvantage, however, is that the cut-off point for LOQs introduces a subjective element in the calculations.

The outcome of these calculations are also indicated in Table 1. For BDE 209 in eel there is only one LOQ that satisfies the criterion for inclusion. This observation, number 1, exhibits a small overlap with the numerical data (observations 9–12, Fig. 3), so that the means of the calculations with and without this LOQ differ little. However, the inclusion of the LOQs for BDE 209 in porpoise liver and for BDE 75 in sediment seven has a pronounced effect on the outcome of the calculations. In each case it is essential to use the calculated

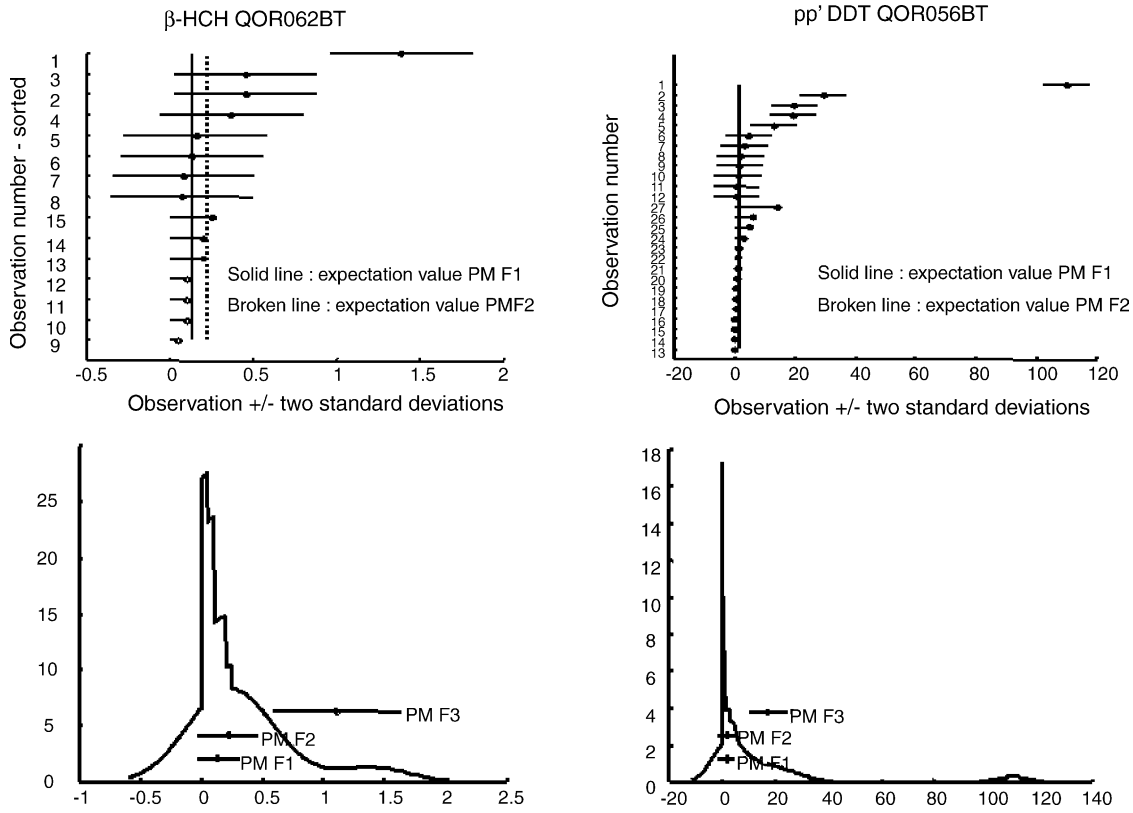

Fig. 3. Overview of results and the summed measurement functions for β-HCH and pp′-DDT in biological tissue showing the need for inclusion of the left-censored values. Expectation value and standard deviation of PMF_i are indicated by horizontal bars in the bottom panels. The β-HCH dataset contains 7 left-censored data (observation numbers 9–15, the pp′-DDT dataset contains 15 left-censored data (observation numbers 13–27).

Table 2
Results of calculations on data from Quasimeme interlaboratory scheme

Quasimeme round	Determinand	Dataset	Nobs	Rectangular pdf			Triangular pdf		
				Mean PMF1	Standard PMF1	% PMF1	Mean PMF1	Standard PMF1	% PMF1
QOR070BT	HCB	All data	35	0.08	0.04	64.9	0.08	0.03	64.9
		Only numerical data	26	0.09	0.03	68.6	0.09	0.03	68.6
		Numerical data and LOQ < limit	27	0.09	0.03	66.5	0.09	0.03	66.5
QOR070BT	pp′-DDE	All data	36	1.31	0.37	63.8	1.31	0.37	64.0
		Only numerical data	34	1.32	0.36	67.2	1.32	0.36	67.2
		Numerical data and LOQ < limit	36	1.31	0.37	63.8	1.31	0.37	64.0
QOR068BT	CB52	All data	31	0.13	0.09	63.7	0.13	0.08	63.2
		Only numerical data	27	0.14	0.08	70.2	0.14	0.08	70.2
		Numerical data and LOQ < limit	30	0.13	0.08	65.2	0.13	0.08	65.2
QOR068BT	CB156	All data	24	0.05	0.04	64.5	0.06	0.04	61.9
		Only numerical data	16	0.06	0.04	70.3	0.06	0.04	70.3
		Numerical data and LOQ < limit	19	0.05	0.04	63.8	0.06	0.04	63.6
QOR068BT	CB180	All data	31	0.18	0.07	69.3	0.18	0.07	68.9
		Only numerical data	29	0.18	0.07	73.5	0.18	0.07	73.5
		Numerical data and LOQ<limit	30	0.18	0.07	71.1	0.18	0.07	71.1
QOR062BT	CB28	All data	29	0.30	0.08	63.8	0.30	0.08	64.1
		Only numerical data	26	0.30	0.08	67.2	0.30	0.08	67.2
		Numerical data and LOQ < limit	27	0.30	0.08	65.3	0.30	0.08	65.5
QOR062BT	pp′-DDT	All data	24	0.14	0.19	53.6	0.17	0.21	48.2
		Only numerical data	13	0.16	0.27	60.0	0.16	0.27	60.0
		Numerical data and LOQ < limit	17	0.12	0.23	51.0	0.13	0.24	50.4
QOR062BT	β-HCH	All data	15	0.13	0.18	57.8	0.16	0.19	54.6
		Only numerical data	8	0.24	0.25	77.2	0.24	0.25	77.2
		Numerical data and LOQ < limit	14	0.14	0.19	57.2	0.17	0.20	54.8
QOR062BT	γ-HCH	All data	23	0.13	0.16	66.8	0.14	0.16	65.6
		Only numerical data	17	0.15	0.18	77.9	0.15	0.18	77.9
		Numerical data and LOQ < limit	21	0.14	0.17	67.4	0.14	0.17	66.9
QTM053BT	Silver	All data	15	15.4	7.2	51.4	15.4	5.7	48.9
		Only numerical data	11	14.8	2.7	62.3	14.8	2.7	62.3
		Numerical data and LOQ < limit	12	14.7	2.8	57.3	14.7	2.8	57.4
QTM054BT	Cadmium	All data	40	6.35	2.67	60.9	6.44	2.48	59.2
		Only numerical data	35	6.30	2.28	62.3	6.30	2.28	62.3
		Numerical data and LOQ < limit	35	6.30	2.28	62.3	6.30	2.28	62.3
QTM054BT	Nickel	All data	29	45.5	34.5	63.7	47.0	33.6	61.6
		Only numerical data	22	48.7	35.1	69.2	48.7	35.1	69.2
		Numerical data and LOQ < limit	23	48.6	35.1	66.2	48.6	35.1	66.2
QTM051BT	Cadmium	All data	30	4.98	3.23	67.9	5.36	3.13	65.0
		Only numerical data	21	4.81	3.28	72.2	4.81	3.28	72.2
		Numerical data and LOQ < limit	23	4.80	3.28	65.9	4.80	3.28	65.9
QTM052BT	Chromium	All data	29	259.5	74.5	65.7	259.6	73.6	65.5
		Only numerical data	27	262.1	70.6	67.7	262.1	70.6	67.7
		Numerical data and LOQ < limit	28	258.6	71.6	66.4	257.9	71.2	66.8
QTM047BT	Silver	All data	11	4.95	5.58	56.3	5.83	5.97	51.8
		Only numerical data	6	6.05	7.81	63.9	6.05	7.81	63.9
		Numerical data and LOQ < limit	9	4.25	6.07	53.1	4.68	6.21	51.7
QOR056BT	op′-DDT	All data	21	1.10	0.96	49.5	1.29	0.93	43.0
		Only numerical data	9	1.25	0.81	64.3	1.25	0.81	64.3
		Numerical data and LOQ < limit	15	0.76	0.83	44.4	0.89	0.84	43.0
QOR056BT	pp′-DDT	All data	27	1.66	3.00	40.3	2.36	3.62	35.2
		Only numerical data	12	2.59	4.57	58.4	2.59	4.57	58.4
		Numerical data and LOQ < limit	23	1.05	2.88	38.6	1.71	3.68	35.0

Table 2 (*Continued*)

Quasimeme round	Determinand	Dataset	Nobs	Rectangular pdf			Triangular pdf		
				Mean PMF1	Standard PMF1	% PMF1	Mean PMF1	Standard PMF1	% PMF1
QOR057BT	CB28	All data	31	0.64	0.40	70.1	0.66	0.38	67.8
		Only numerical data	24	0.67	0.38	79.0	0.67	0.38	79.0
		Numerical data and LOQ < limit	27	0.63	0.38	72.6	0.64	0.37	72.7
QOR057BT	CB52	All data	32	1.09	0.59	69.1	1.11	0.57	66.5
		Only numerical data	26	1.08	0.53	76.2	1.08	0.53	76.2
		Numerical data and LOQ < limit	28	1.07	0.54	71.2	1.07	0.54	71.2
QOR057BT	CB101	All data	32	2.17	0.64	73.1	2.18	0.64	73.1
		Only numerical data	30	2.19	0.62	75.0	2.19	0.62	75.0
		Numerical data and LOQ < limit	31	2.17	0.62	73.4	2.16	0.62	73.7
QOR057BT	CB105	All data	29	0.55	0.27	66.3	0.56	0.26	63.9
		Only numerical data	23	0.55	0.24	71.9	0.55	0.24	71.9
		Numerical data and LOQ < limit	25	0.54	0.24	67.7	0.54	0.24	67.9

values along with expert judgement on the whole information base.

5. Case study II: examples from the QUASIMEME laboratory performance studies of determinands in marine matrices

QUASIMEME[3] is an international interlaboratory scheme supporting the quality assurance of environmental measurements in the marine environment [15]. As such this scheme regularly encounters datasets which contain *less than* values. The results of calculations on a selected number of such datasets are given in Table 2.

The proportion of *less than* values in relation to the number of numerical data ranges from 5 to 133% for all LOQs and from 0 to 75% for the LOQs which satisfy the criterion 'less than the mean of dataset of numerical data'. Inspection of the results of the calculations (Table 2) confirm that inclusion of *less than* values in the calculations has, in the majority of cases, only a small effect on the expectation value in spite of the relatively large proportions of *less than* values. Exceptions are provided by the datasets β-HCH in QOR062BT, silver in QTM047BT, op′-DDT in QOR056BT and op′-DDT in QOR056BT. These three cases are small datasets with 80–130% LOQs relative to numerical data. Inclusion of the *less than* values gives rise to a significant lowering of the expectation value. The datasets β-HCH and pp′-DDT are discussed into more detail and presented in Fig. 3. Many organochlorines occur at low concentrations, at or below the limit of quantification in lean fish tissue. Often in such analyses the numerical data may be a result of contamination during sampling or subsequent chemical analysis.

Of the fifteen values in the data set for β-HCH, seven were left-censored values. The overview of results clearly shows the overlap of the left-censored data with most of the numerical values and the interactions of the pdfs of both the normal distribution and the rectangular distributions (Fig. 3). A similar situation occurs with the pp′DDT. Of the 27 values, 15 are left-censored. The left-censored data interact strongly with all but about five of the numerical data. These five numerical data stand out in the dataset by their high values and have a negligible interaction with the other data. This is also reflected in pattern of the overall measurement function for pp′-DDT (Fig. 3).

These two examples clearly illustrate the importance of being able to include the left-censored values in the overall assessment of laboratory performance studies where natural unspiked materials are used and the concentrations of the determinands are close to the LOQs.

In such situations the LOQs may more realistically represent a closer estimate of the true concentration and should be included in the assessment.

6. Case study III—temporal trend of dissolved cadmium in the rhine river at Lobith

The Dutch Directorate General for Public Works and Water Management is responsible for the monitoring of water quality in the river Rhine. The water quality has improved significantly in the past decade. As a result, the number of *less than* values has increased as concentrations have decreased below the performance characteristics of the methodology employed. In Fig. 4, the temporal trend of dissolved cadmium in Rhine water at Lobith is given for the period between 1988 and 1998. In the first part of this period, *less than* values were hardly observed. In the second half of this decade, *less than* values constituted up to 75% of the data in a year. Neglecting these *less than* values gives a false, high assessment of the dissolved cadmium concentrations, since from about 1996

[3] QUASIMEME: quality assurance of information for marine environmental monitoring.

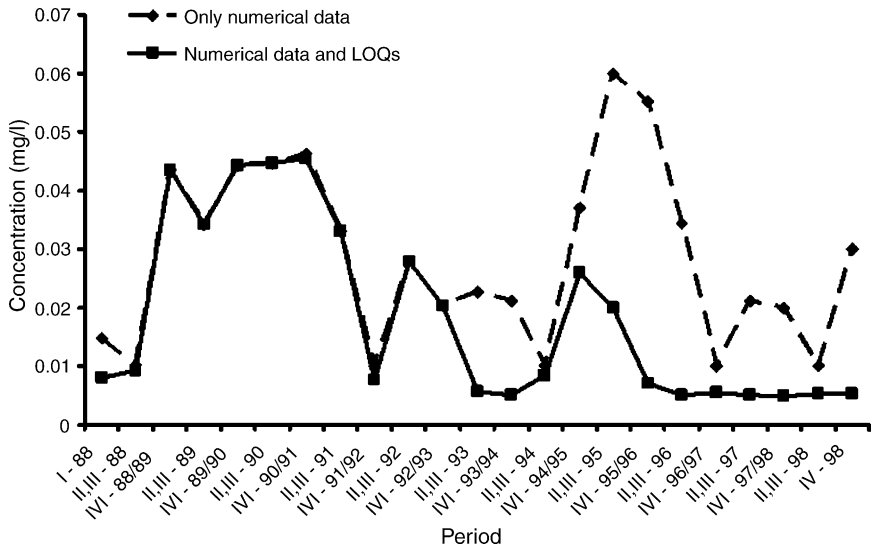

Fig. 4. Temporal trend of dissolved cadmium in the Rhine river at Lobith between 1988 and 1999. Sampling frequency was 12 times/year in the first 5 years and 24 times/year in the last 5 years. Limit of detection was 0.01 or 0.02 mg/l. A rectangular pdf was used to include the LOQs into the calculations.

onwards the *less than* values dominate the dataset (Fig. 4). Therefore inclusion of the *less than* values becomes critical to the reporting of this type of monitoring activity and any subsequent environmental control policies.

7. Conclusions

In this paper the inference model from [12] is extended to include left-censored data into the assessment. The model associates a measurement function associated with each observation. A measurement function is obtained as the square root of the probability density function proper. Left-censored data are included in the model by establishing the appropriate pdfs and constructing the corresponding measurement functions. As such, *less than* data and numerical values are treated in the same manner. Consequently, there are no limits to the magnitudes of the LOQs nor to the number of LOQs in the dataset.

The results of the model have been compared with those obtained using the Cohen maximum likelihood estimator for simulated datasets with single-valued LOQs. The agreement between the two methods depends on the number of LOQs present in the dataset. When the number of LOQs is high, the model may give a principally different outcome. Conventionally interpreted, the model may designate the numerical values as outliers where many LOQs are present in the dataset. On the contrary, the Cohen maximum likelihood estimator always has to infer the mean and standard deviation of the dataset from the numerical data, adjusting for the LOQs.

The model has been applied to datasets obtained from interlaboratory studies and from water quality monitoring. In specific cases, the presence of LOQs with values greater than the mean of the numerical data may contribute significantly to the outcome of the calculations. The indicative information of such LOQs may be very limited. This effect can be circumvented by imposing criteria on the acceptance of the LOQs before inclusion in the calculations, although this approach also introduces an element of subjectivity. Expert judgement is required in such cases. The calculations demonstrate that the model can handle *less than* values well while retaining its features of graphical and quantitative output. For each mode or cluster of data the model provides the respective expectation value, standard deviation and percentage of data represented by the mode in the dataset. The graphical output provides an overview of the key features of the structure of the data that allows further exploratory analysis and an informed evaluation.

The calculations presented in this paper confirm that inclusion of left-censored data in the calculations of population characteristics improves an assessment procedure.

References

[1] W.M. Daniels, N.A. Higgins, Environmental distributions and the practical utilization of detection limited environment measurement data, National Radiological Protection Board, 2002, ISBN 0859514846.
[2] S. Kuttatharmmakul, D.L. Massart, D. Coomans, J. Smeyers-Verbeke, Anal. Chim. Acta 441 (2001) 215–229.
[3] A. Singh, J. Nocerino, Chemom. Intell. Lab. Syst. 60 (2002) 69–86.
[4] US Army Engineer Waterways Experiment Station, Guidelines for Statistical Treatment of Less Than Detection Limit Data in Dredged Sediment Evaluations, Environmental Effects of Dredging, Technical Notes EEDP-04-23, 1995.
[5] J. de Boer, W.P. Cofino, Chemosphere 46 (2002) 625–633.
[6] D.C. Glass, C.N. Gray, Ann. Occup. Hyg. 45 (2001) 275–282.
[7] J.F. England, R.D. Jarrett, J.D. Salas, J. Hydrol. 278 (2003) 172–196.

[8] R.J. Gilliom, D.R. Helsel, Water Resour. Res. 22 (1986) 135–146.
[9] D.R. Helsel, R.J. Gilliom, Water Resour. Res. 22 (1986) 147–155.
[10] D.R. Helsel, T.A. Cohn, Water Resour. Res. 24 (1988) 1997–2004.
[11] O.R. Gilbert, Statistical Methods for Environmental Pollution Monitoring, Wiley, NY, 1987, ISBN 0-471-28878-0.
[12] W.P. Cofino, I. van Stokkum, D.E. Wells, R.A.L. Peerboom, F. Ariese, Chemom. Intell. Lab. Syst. 53 (2000) 37–55.
[13] D. Montville, E. Voigtman, Talanta 59 (2003) 461–476.
[14] P.J. Lowthian, M. Thompson, Analyst 127 (2002) 1359–1364.
[15] D.E. Wells, W.P. Cofino, Mar. Pollut. Bull. 35 (1997) 146–155.