ELSEVIER

Journal of Chromatography A, 975 (2002) 123–133

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

How to estimate moments and quantiles of environmental data sets with non-detected observations? A case study on volatile organic compounds in marine water samples

Tom Huybrechts[a], Olivier Thas[b], Jo Dewulf[a], Herman Van Langenhove[a,]*

[a]*Department of Organic Chemistry, Faculty of Agricultural and Applied Biological Sciences,*
Ghent University, Coupure Links 653, B-9000 Ghent, Belgium

[b]*Department of Applied Mathematics, Biometrics and Process Control, Faculty of Agricultural and Applied Biological Sciences,*
Ghent University, Coupure Links 653, B-9000 Ghent, Belgium

Abstract

Concentrations of 27 priority volatile organic compounds were measured in water samples of the North Sea and Scheldt estuary during a 3-year monitoring study. Despite the use of a sensitive analytical method, a number of data were censored. That is, some concentrations were below the decision limit or critical level defined by IUPAC. To characterize the observed measurement results, an attempt was made to identify an appropriate procedure to compute summary statistics for the censored data sets. Several parametric and robust parametric approaches based on the maximum likelihood principle and probability-plot regression method were evaluated for the estimation of the mean, standard deviation, median and interquartile range using three uncensored analytes (1,1,2-trichloroethane, tetrachloroethene and *o*-xylene) from the monitoring survey. Performance was assessed by artificially censoring the observed concentrations and estimating moments and quantiles at each censoring level. Results showed that methods with the least distributional assumptions, such as the robust bias-corrected restricted maximum likelihood method, perform best for estimating the mean and standard deviation, while both parametric and robust parametric techniques can be used for quantiles. Hence, summary statistics could be estimated with little bias (5–10%) up to 80% of censoring for the data sets employed in this study.

Keywords: Mean; Median; Water analysis; Standard deviation; Maximum likelihood estimation; Probability-plot regression; Decision limit; Detection limit; Censored data; Interquartile range; Robust imputation; Volatile organic compounds

1. Introduction

Due to their widespread occurrence and fate in the marine environment, a number of chlorinated short-chain hydrocarbons (CHCs), monocyclic aromatic hydrocarbons (MAHs) and chlorinated monocyclic aromatic hydrocarbons (CMAHs) were classified as ‘‘priority’’ and ‘‘priority toxic’’ pollutants at the 3rd International Conference for the Protection of the North Sea [1]. Recently, several volatile organic compounds (VOCs), for the most part CHCs, MAHs and CMAHs, were proposed by the Marine Chemistry Working Group for inclusion in the EU Water Framework Directive 2000/60/EC [2], and some CMAHs were listed by the OSPAR (Oslo and Paris) Commission as Chemicals for Priority Action [3]. Although of major environmental concern, VOCs

*Corresponding author. Tel.: +32-9-264-5953; fax: +32-9-264-6243.

E-mail address: herman.vanlangenhove@rug.ac.be (H. Van Langenhove).

0021-9673/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0021-9673(02)01327-4

have received much less attention in marine pollution research compared to other priority pollutants.

To acquire data on concentration levels of 27 VOCs, surface waters were sampled in the North Sea and Scheldt estuary during a 3-year monitoring study, and analysed by purge-and-trap/gas chromatography–mass spectrometry [4]. Analytes included chlorinated C_1–C_4 alkanes and alkenes, MAHs and CMAHs, all selected from formerly published priority lists [1].

Target VOCs were commonly found at trace level concentrations in marine and estuarine waters, typically ng l^{-1}. Despite the use of a sensitive analytical method and state-of-the-art laboratory instrumentation, measurement results were attended by the problem of detection and, as we will see, of censoring.

IUPAC considers three limiting levels to describe the detection capability of an analytical method: (i) the decision limit or critical level ‘‘at which one may decide whether or not the result of an analysis indicates detection’’, (ii) the detection limit ‘‘at which a given analytical procedure may be relied upon to lead to detection’’ and (iii) the quantification limit ‘‘at which a given procedure will be sufficiently precise to yield a satisfactory quantitative estimate’’ [5]. Measured values below the decision limit cannot be distinguished from sample blanks and should be reported as ‘‘less than the detection limit’’ rather than as numerical values. These data points are referred to as ‘‘censored at the decision limit’’. Censoring at the detection or even at the quantification limit would lead to unnecessary loss of information since numerical values are available.

Censored data points were commonly found for most target compounds in the monitoring survey. Statistically, these data are designated as ‘‘left-censored’’ since the left or lower portion of the distribution cannot be observed. Two types of censoring are usually considered. The situation where all data below a fixed value are censored is called type I censoring. With type I censoring, the number of values censored is a random variable. In type II censored data sets, a fixed number of data points are always censored and the censoring threshold is a random variable. Censored water quality data should resemble the first type because the censoring threshold is fixed by the analytical method.

The occurrence of censored measurements greatly complicates statistical analysis of environmental data. Standard calculation methods fail as only part of the data points are numerically known, while the other fraction is only known to occur within a restricted range of values. However, it is often of particular interest to estimate moments and quantiles for the purpose of characterizing the observed data.

Discarding censored measurements from calculations should not be considered since it involves loss of important information and yields biased estimates.

Replacement techniques are most frequently used by environmental and analytical chemists. These methods substitute a constant value such as 0, the censoring limit or one half the censoring limit for each ‘‘less than’’ data point. A ‘‘complete’’ data set is obtained and standard calculation methods can be used. Although frequently applied for ease of computation, these methods have no theoretical basis. Furthermore, simple replacement techniques perform poorly in comparison to statistical methods [6–9].

The use of statistically sound procedures often results in less biased estimates and are therefore highly recommended. The most commonly employed statistical methods have been reviewed by Helsel and Hirsch [10]. While such techniques are well established within the statistical community, they are not as well known by chemists and scientists involved in environmental or analytical studies.

For the purpose of identifying an appropriate method to estimate summary statistics for the censored data sets observed in the survey, several statistical methods were selected from literature based on results of previous simulation studies [6–17]. Their performance was checked using actual uncensored data sets from the monitoring study. Statistics of interest included the mean, standard deviation, median and interquartile range.

2. Materials and methods

2.1. General

Surface waters were sampled twice a year in the Channel, the Belgian continental shelf, the southern North Sea and the Scheldt estuary during a 3-year monitoring study (April 1998 to October 2000) to determine concentration levels of 27 priority VOCs. The analytes of interest included 1,1-dichloroethene (DCE11), dichloromethane (CH_2Cl_2), *trans*-1,2-dichloroethene (DCE12), 1,1-dichloroethane (DCA11), chloroform ($CHCl_3$), 1,1,1-trichloroethane (TRI111), cyclohexane (CYCLO), tetrachloromethane (CCl_4), 1,2-dichloroethane (DCA12), benzene (BENZ), trichloroethene (TCE), 1,2-dichloropropane (DCP12), toluene (TOL), 1,1,2-trichloroethane (TRI112), tetrachloroethene (PCE), chlorobenzene (ClBENZ), ethylbenzene (EtBENZ), *m*- and *p*-xylene (MPXYL), *o*-xylene (OXYL), 1,3-dichlorobenzene (DCB13), 1,4-dichlorobenzene (DCB14), 1,2-dichlorobenzene (DCB12), 1,3,5-trichlorobenzene (TCB135), 1,2,4-trichlorobenzene (TCB124), hexachloro-1,3-butadiene (HCB) and 1,2,3-trichlorobenzene (TCB123). As a result, 47 marine and 84 estuarine water samples were analysed by purge-and-trap combined with high resolution gas chromatography and detection by mass spectrometry operating in the selected ion monitoring (SIM) mode. Data were produced by analyses deemed ‘‘in control’’ by a rigorous quality assurance/quality control (QA/QC) program [4]. Measurements were grouped into two multivariate data sets, one labeled as ‘‘North Sea’’, the other as ‘‘Scheldt estuary’’. Each variable was censored at the decision limit and reported as ‘‘$<x_{\text{det}}$’’, with x_{det} the numerical value of the detection limit. A single censoring value was used for each compound throughout the study. Censoring intensities for each analyte in both data sets are given in Fig. 1.

2.2. Approach

Three compounds remained uncensored: 1,1,2-trichloroethane and tetrachloroethene (‘‘Scheldt estuary’’), and *o*-xylene (‘‘North Sea’’). Each of these data sets was artificially censored by gradually discarding the lowest uncensored measurement from the ranked set of values. At each degree of censoring, the mean, standard deviation (SD), median and interquartile range (IQR) were estimated by several statistical procedures described below. Method performance was assessed by a relative error- or bias-term, calculated as $100\big((\hat{\theta}_{\text{censor}} - \hat{\theta}_{\text{uncensor}})/\hat{\theta}_{\text{uncensor}}\big)$,

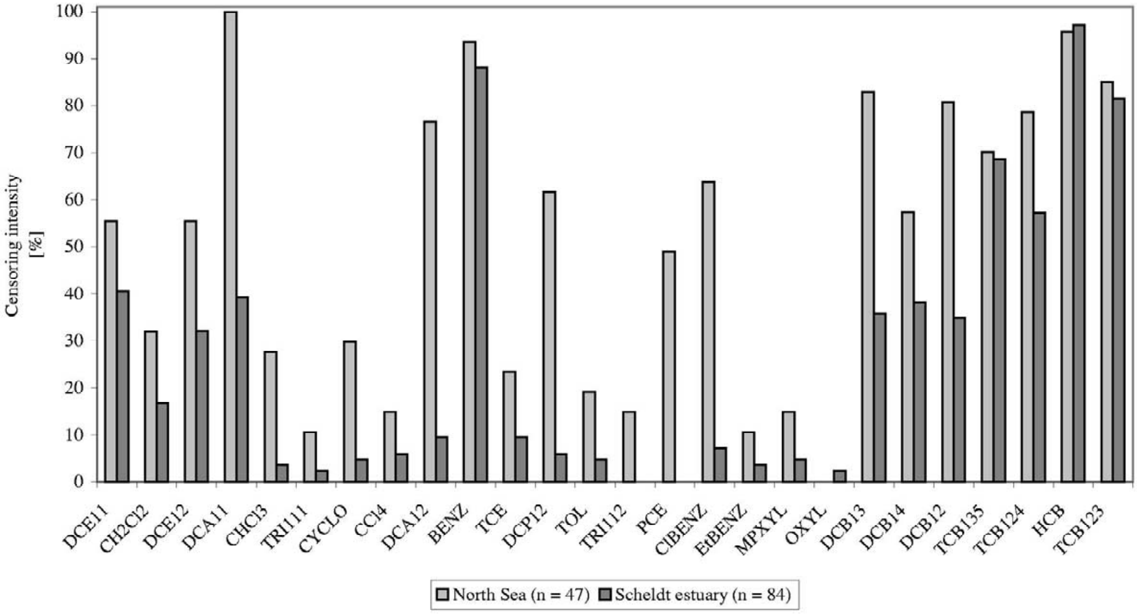

Fig. 1. Censoring intensities for 27 priority VOCs measured in the North Sea ($n = 47$) and Scheldt estuary ($n = 84$, except trichlorobenzenes and hexachloro-1,3-butadiene $n = 70$, and 1,2-dichlorobenzene $n = 83$). The abbreviations for each analyte are given in the Materials and methods. Since *m*- and *p*-xylene could not be separated, they are reported as a single value (MPXYL).

with $\hat{\theta}_{\text{censor}}$ and $\hat{\theta}_{\text{uncensor}}$ sample parameter estimates computed from censored and uncensored data sets, respectively. A hypothetical decision limit was calculated at each censoring level as the average of the highest censored and the lowest uncensored measurement. This went on until 90% of the raw data was censored.

2.3. Estimation methods

2.3.1. Parametric methods

Parametric or distributional methods use the characteristics of an assumed underlying distribution to estimate moments and quantiles. Given a distribution, estimates of summary statistics are computed that best match the observed concentrations above the decision limit and the percentage of data below the limit. Parametric methods have originally been derived for normally distributed variables. They can be employed if the underlying distribution is not normal but can be transformed to normal form. Environmental data are often positively skewed and the lognormal distribution, which is a simple transform of a normal distribution, is commonly used as a model.

2.3.1.1. Cohen's maximum likelihood method

Consider an ordered data set $x_1 \leq x_2 \ldots \leq x_c \leq x_{c+1} \ldots \leq x_n$, where the first c observations out of n measurements are censored. Assume that the variable x_i can be described adequately by a lognormal distribution. Let ln $(x_i) = y_i$ for $i = c+1, \ldots, n$ and let y_{censor} be the natural logarithm (ln) of the decision or censoring limit. The likelihood function L for the data is given by:

$$L(\mu_y, \sigma_y) = \frac{n!}{(n-c)!c!}\left(\Phi\left(\frac{y_{\text{censor}} - \mu_y}{\sigma_y}\right)\right)^c \frac{1}{\sqrt{\left(2\pi\sigma_y^2\right)^n}} \exp\left(\frac{\sum_{i=1}^{n}(y_i - \mu_y)^2}{-2\sigma_y^2}\right) \tag{1}$$

with Φ the cumulative distribution function of a standard normal variate, μ_y the mean and σ_y the standard deviation of the ln-transformed data.

The maximum likelihood estimates, $\hat{\mu}_y$ and $\hat{\sigma}_y$, of μ_y and σ_y can be found by calculating the values of μ_y and σ_y that maximize the function L. By taking the natural logarithm of (1) and setting the partial derivatives with respect to μ_y and σ_y to zero, the maximum likelihood estimators $\hat{\mu}_y$ and $\hat{\sigma}_y$ can be calculated.

Cohen [11] proposed the following solution:

$$\hat{\mu}_y = m_y - \lambda(m_y - y_{\text{censor}}) \tag{2}$$

$$\hat{\sigma}_y = \sqrt{s_y^2 + \lambda(m_y - y_{\text{censor}})^2} \tag{3}$$

with m_y and s_y the sample mean and standard deviation of the $n - c$ uncensored ln-transformed observations, respectively. The factor λ is calculated from the proportion of censored data $h = c/n$ and $\hat{\gamma} = s_y^2/(m_y - y_{\text{censor}})^2$. Tables are provided to determine λ from these parameters [11]. We used a power series expansion by Haas and Scheff [8] that fits the tabulated values to within 6% relative error.

2.3.1.2. Bias-corrected restricted maximum likelihood method

Although less frequently considered for calculating summary statistics from censored environmental data, the one-step restricted maximum likelihood estimators are somewhat simpler to compute [7,8,14]. The method provides the following explicit solution to maximize Eq. (1) for the mean and standard deviation by imposing an assumption that the number of observations below the censoring limit follow a binomial distribution.

Estimators of the mean and standard deviation for censored normally distributed data y_i are given by:

$$\hat{\mu}_y = y_{\text{censor}} - \Phi^{-1}(h)\hat{\sigma}_y \tag{4}$$

$$\hat{\sigma}_y = 0.5\left(\frac{a\Phi^{-1}(h)}{n-c} + \sqrt{\left(\frac{a\Phi^{-1}(h)}{n-c}\right)^2 + 4\frac{b}{n-c}}\right) \tag{5}$$

where $\Phi^{-1}(h)$ is the inverse cumulative normal distribution function evaluated at h, the proportion of censored data. The parameters a and b are calculated as:

$$a = \sum_{i=c+1}^{n}(y_i - y_{\text{censor}})$$

and

$$b = \sum_{i=c+1}^{n} (y_i - y_{\mathrm{censor}})^2$$

Since the estimators are not asymptotically unbiased at low degrees of censoring, the following bias-corrected estimators of the mean and standard deviation were suggested by Haas and Scheff [8],

$$\hat{\mu}_{y,bc} = \frac{a'}{n-c} - \xi\hat{\sigma}_y \tag{6}$$

$$\hat{\sigma}_{y,bc} = \sqrt{\frac{b'}{n-c} - \left(\frac{a'}{n-c}\right)^2 - \left(\xi\Phi^{-1}(h) - \xi^2\right)\hat{\sigma}_y^2} \tag{7}$$

where the parameters a' and b' are calculated by $a' = \Sigma_{i=c+1}^{n} y_i$ and $b' = \Sigma_{i=c+1}^{n} y_i^2$, and the correction term ξ is given by:

$$\xi = \frac{n}{(n-c)\sqrt{2\pi}} \exp\left(-0.5\left(\Phi^{-1}(h)\right)^2\right)$$

2.3.1.3. Probability-plot regression method

It is possible to estimate the mean and standard deviation of a censored lognormally distributed data set based on a linear relationship of the ln-transformed uncensored values versus the normal scores z_i. The latter are designated by the plotting positions p_i of the ordered uncensored measurements with $z_i = \Phi^{-1}(p_i)$ and $\Phi^{-1}(p_i)$ the inverse cumulative normal distribution function evaluated at p_i [13]. Again, consider that the first c out of n observations are censored. Plotting positions p_i for the uncensored observations are given by the Hirsch–Stedinger Blom-based equation [17]:

$$p_i = \frac{c}{n} + \frac{n-c}{n}\left(\frac{i - 0.375 - c}{n + 0.25 - c}\right) \quad i = c+1, \ldots, n \tag{8}$$

where i is the rank of the ith measurement.

The mean $\hat{\mu}_y$ and standard deviation $\hat{\sigma}_y$ are estimated by ordinary least squares as the intercept and regression coefficient, respectively, in a simple linear regression model of the ln-transformed uncensored values against the corresponding normal scores.

2.3.1.4. Calculation of summary statistics

The methods described above estimate the mean μ_y and standard deviation σ_y of ln-transformed data y_i. To compute the estimated mean $\hat{\mu}_x$ and standard deviation $\hat{\sigma}_x$ of the original lognormal data set x_i, the following back-transformation is required:

$$\hat{\mu}_x = \hat{\mu}_y + 0.5\hat{\sigma}_y^2 \tag{9}$$

$$\hat{\sigma}_x = \sqrt{\hat{\mu}_x^2\left(\exp\left(\hat{\sigma}_y^2\right) - 1\right)} \tag{10}$$

Besides the estimation of the mean and standard deviation, the median and interquartile range are of interest. The interquartile range is calculated as the difference between the 75th percentile and the 25th percentile, and represents the range of the central 50% of the data.

To estimate quantiles, the following equation was used:

$$\hat{x}_q = \exp\left(\hat{\mu}_y + \Phi^{-1}(q)\hat{\sigma}_y\right) \tag{11}$$

where $\hat{x}_q$ is an estimate of x_q, a quantile with non-exceedence probability of q percentage and $\Phi^{-1}(q)$ is the inverse of the standard normal cumulative distribution function evaluated at the qth percentile.

2.3.2. Robust parametric methods

A distribution is fit to the data but unlike previous methods, the fitted distribution is used only to obtain values for the c observations below the censoring limit. These extrapolated values, denoted by $x_i^{(s)}$, are not considered estimates of censored observations, but are used collectively with uncensored measurements to compute summary statistics.

The regression relationship defined in the previous paragraph is used to assign plotting positions to the c censored measurements:

$$p_i = \frac{c}{n}\left(\frac{i - 0.375}{c + 0.25}\right) \quad i = 1, \ldots, c \tag{12}$$

The censored observations are then computed by:

$$x_i^{(s)} = \exp\left(\hat{\mu}_y + \hat{\sigma}_y\Phi^{-1}(p_i)\right) \quad i = 1, \ldots, c \tag{13}$$

with $\hat{\mu}_y$ the mean and $\hat{\sigma}_y$ the standard deviation estimated from the ln-transformed data using one of the methods described above.

A ‘‘complete’’ data set is obtained and the mean and standard deviation can be estimated by the method of moments. Several authors used this fill-in or imputation technique only in combination with the probability-plot regression method [6,9,12,17]. However, any parametric method can be used as shown by El-Shaarawi [14], and Kroll and Stedinger [17].

3. Results and discussion

Table 1 summarizes the characteristics of the three uncensored data sets. A large difference between the mean and median suggests some degree of skewness towards lower concentration levels.

To check the validity of the lognormal distribution, the Kolmogorov–Smirnov method with Lillifors significance correction was applied to the ln-transformed observations. The results are listed in Table 2. At a significance level of $\alpha = 0.05$ only *o*-xylene can be approximated by a lognormal distribution. Lack of lognormality for 1,1,2-trichloroethane and tetrachloroethene results from the presence of large outliers in the right tail of the distribution as can be seen in Fig. 2. Outliers are sometimes discarded to fit the remaining observations to a lognormal distribution. However, deletion of outliers on the basis of statistical grounds should be avoided. They are often interesting results and should be investigated further.

While assumptions regarding the underlying distribution of uncensored data sets are relatively easy to check, things get more complicated when censored measurements are present. As the degree of censoring increases, the information available to establish the nature of the distributional shape decreases. Nevertheless, it has become a standard practice to perform statistical analyses of censored data sets in the logarithmic scale. To assess the effect of non-lognormality on lognormal-based statistics, we will continue to assume that all observations fit a lognormal distribution and use ln-transformed data for the purpose of calculations.

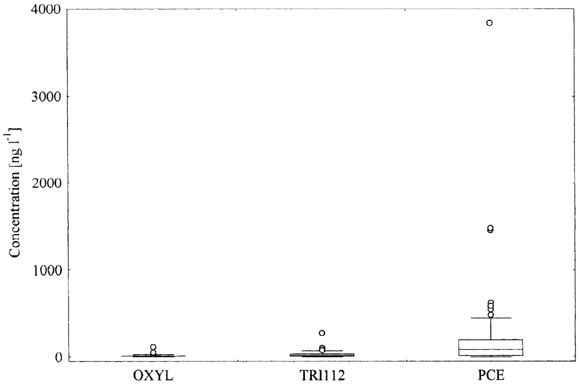

Fig. 2. Box-whisker plots for *o*-xylene ($n = 47$), 1,1,2-trichloroethane and tetrachloroethene ($n = 84$).

3.1. Mean and standard deviation

To limit the number of graphs, we will only show results of *o*-xylene and tetrachloroethene. Although tetrachloroethene and 1,1,2-trichloroethane do not display the exact same pattern of estimates, findings

Table 1
Summary statistics of raw data sets (ng l^{-1})

Compound	Sampling location	*n*	Mean	SD	Median	IQR
o-Xylene	North Sea	47	12.9	17.8	6.90	11.1
1,1,2-Trichloroethane	Scheldt estuary	84	26.7	35.0	16.1	27.2
Tetrachloroethene	Scheldt estuary	84	199	473	84.5	181

n, sample size; SD, standard deviation; IQR, interquartile range.

Table 2
Results of the Kolmogorov–Smirnov test with Lillifors significance correction (KS) for uncensored ln-transformed observations

Compound	KS-distance	df	*P*-value	Passed normality test? ($\alpha = 0.05$)
o-Xylene	0.111	47	0.188	Yes
1,1,2-Trichloroethane	0.105	84	0.024	No
Tetrachloroethene	0.118	84	0.006	No

df, degrees of freedom; *P*, probability.

are similar for both compounds. Less bias is observed in the case of 1,1,2-trichloroethane, especially at lower censoring levels.

Fig. 3 shows the performance of Cohen's maximum likelihood (CML), the bias-corrected restricted maximum likelihood (BRML) and probability-plot regression (PPR) method for estimating the mean and standard deviation of artificially censored data sets of *o*-xylene.

Since assumptions regarding the underlying distribution of *o*-xylene are correct, one intuitively expects bias to be zero or at least minimal at the lowest censoring intensities, and increase as more data are artificially censored. That is, the less information is available to establish the true distributional shape, the poorer the estimates. However, results indicate that estimates of the mean and standard deviation are biased, ±6% and 25% respectively, at the lowest censoring level. While relative error then slowly declines as censoring increases, CML estimates of the mean become increasingly biased at censoring levels above 45%. The PPR and BRML estimators, however, show little bias (1–3%) up to 70% and 80% of censoring, respectively. Estimates of the standard deviation exhibit a similar pattern at censoring levels below 45%, with little difference between the various methods. Bias of the CML and PPR estimates increases exponentially,

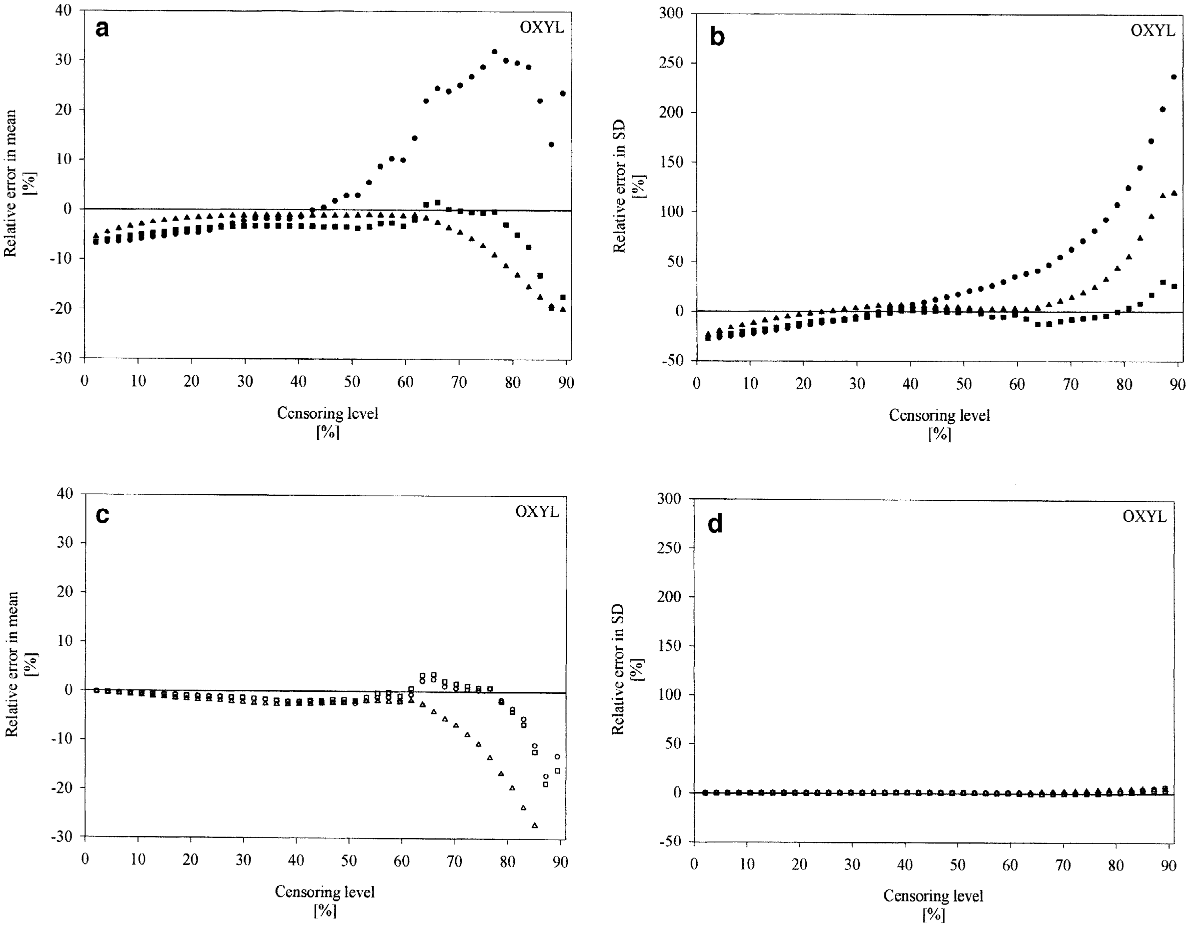

Fig. 3. Comparative performance of statistical methods for estimating the mean and standard deviation (SD) of artificially censored *o*-xylene data. ● Cohen's ML; ■ bias-corrected restricted ML; ▲ probability-plot regression; ○ robust Cohen's ML; □ robust bias-corrected restricted ML; △ robust probability-plot regression.

yielding large relative errors at censoring levels above 45% and 60%, respectively. The BRML method remains less biased at moderate to high censoring intensities.

According to Helsel and Hirsch [10], even if a data set fits the assumed distribution, estimates of moments will be biased from the lowest censoring level on. Bias originates from Eqs. (9) and (10), and is inherent to computing estimates of the mean and standard deviation and then transforming them back to original units. This might explain the occurrence of bias at the lowest censoring intensities for *o*-xylene. An attempt to correct for this "transformation bias" was given by El-Shaarawi [12]. However, estimates of the mean did not improve when this correction factor was used (results not shown).

Besides transformation bias, tetrachloroethene and 1,1,2-trichloroethane fail to assume a lognormal distribution while lognormality was assumed in the calculations. The mean and standard deviation are very sensitive to values of the largest observations and lack of fit of the assumed distribution to these data may result in poor estimates of moments. Results are shown for tetrachloroethene in Fig. 4. All estimators of the mean show high relative errors, 30–40%, at the lowest censoring intensities. Unlike *o*-xylene, large differences are noticed between the three procedures. Relative error decreases at higher

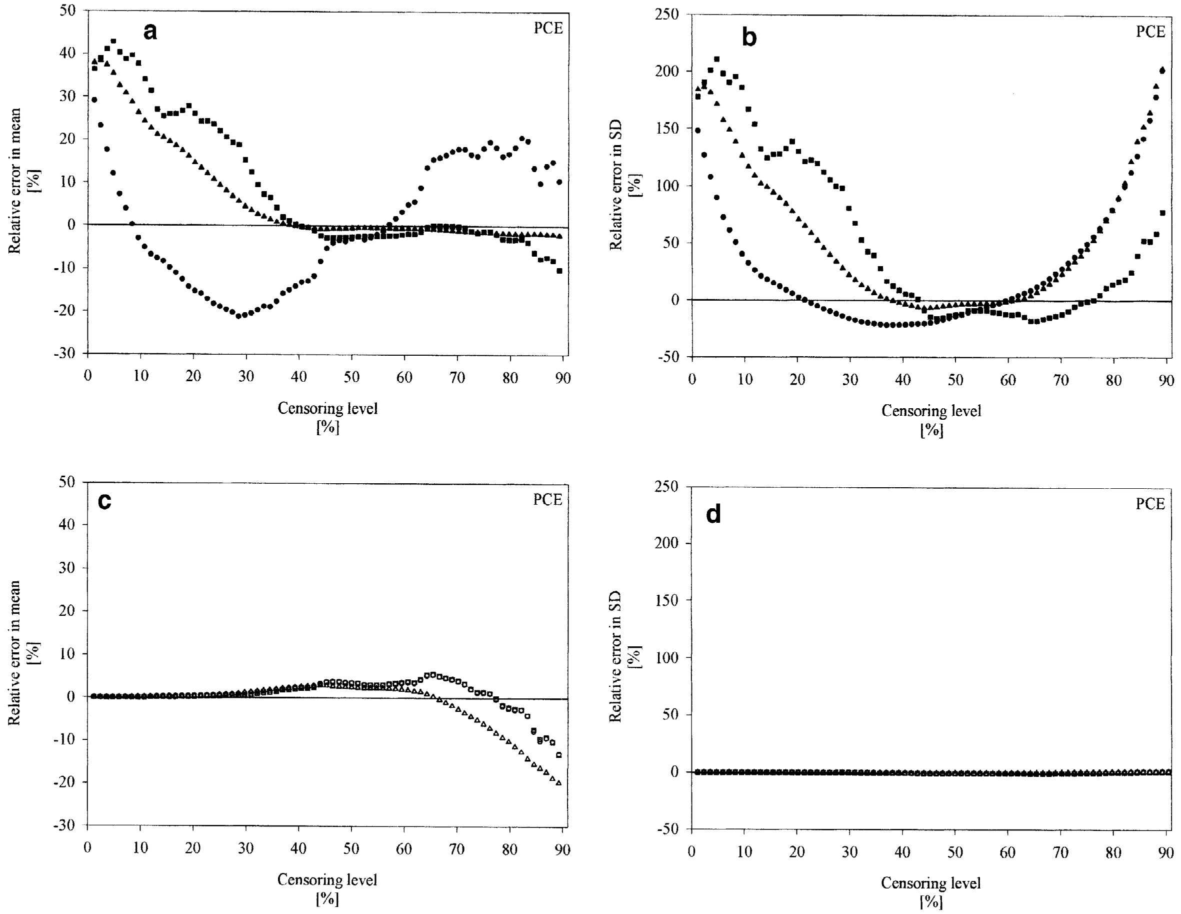

Fig. 4. Comparative performance of statistical methods for estimating the mean and standard deviation (SD) of artificially censored tetrachloroethene data. ● Cohen's ML; ■ bias-corrected restricted ML; ▲ probability-plot regression; ○ robust Cohen's ML; □ robust bias-corrected restricted ML; △ robust probability-plot regression.

censoring levels for the BRML and PPR methods, resulting in minimally biased estimates between 40% and 70–80% of censoring. Cohen's maximum likelihood method yields highly biased estimates over almost the entire censoring range. Estimates of the standard deviation display a similar pattern at censoring levels below 60%, although higher relative errors are observed. At higher censoring intensities bias increases exponentially, especially in the case of CML and PPR estimates, resulting in U-shaped curves with minimum bias around 55% of censoring. The BRML estimator, however, yields less biased estimates at high censoring levels. Relative error of all estimates are much higher compared to the mean.

The implementation of a robust imputation method with each parametric technique significantly enhances precision of estimates of the mean and standard deviation. This is clearly shown for *o*-xylene in Fig. 3. Unlike previous results, all estimates of the mean and standard deviation show very little bias at low censoring levels. Relative error increases only slightly as more data are censored. The three estimators of the mean show similar performance up to 60% of censoring, with bias below 5%. Relative error then becomes increasingly negative for the PPR method, while bias remains within acceptable limits (0–5%), up to 80% of censoring, for the CML and BRML estimates. The latter methods perform equally well. Little difference is noticed between the three methods for estimates of the standard deviation. Bias remains low (0–5%) up to large censoring degrees.

Similar results are obtained for tetrachloroethene (and 1,1,2-trichloroethane) as shown in Fig. 4.

Robust methods are not as sensitive to lack of fit of the largest observations to the assumed distribution because actual observed data are used instead of a fitted distribution above the decision limit. Moreover, estimates of extrapolated values can be directly re-transformed to original units without transformation bias. These methods are referred to as ''robust'' since they perform well even when the data are not lognormally distributed [10].

3.2. Median and interquartile range

As mentioned previously, outliers are commonly found in environmental data sets. Since the mean and standard deviation are inflated by the presence of outlying observations, one can question their ability to adequately describe the central tendency and spread of the data. The median and interquartile range (IQR) are therefore often preferred since they are more robust to outlying measurements. Quantiles have another advantage when applied to censored data: when less than 50% of the data are censored, the sample median is accurately known. Similarly, when less than 25% of the data are below the decision limit, the IQR can be calculated. All censored values can be replaced by any randomly chosen set of values smaller than the lowest detected observation, and the quantiles of interest can be computed by standard calculation methods.

Estimation of the median and interquartile range at censoring levels higher than 50% and 25%, respectively, is shown for tetrachloroethene in Fig. 5.

All three estimators of the median exhibit equally low, negative bias between 50% and 60% of censoring. While relative errors of the PPR estimates become more negative, the CML and BRML estimators generate positively biased values over a short censoring interval before displaying the same downward trend. Little difference is noticed between the maximum likelihood methods over the entire censoring range.

BRML and PPR estimates of the interquartile range show constant bias of 10% from 25 to 60% of censoring. While bias again declines to increasingly negative values for the PPR method, relative error of the BRML estimates remains constant up to 80% of censoring. The CML estimates are highly biased at a censoring level just above 25%. Relative error then decreases, and between 60% and 80% of censoring becomes even smaller than the BRML estimator.

Unlike the estimation of moments, little or no improvement is noticed when the robust imputation method is implemented for computation of the median. Also for the interquartile range, except for Cohen's maximum likelihood method at the low censoring levels, bias is not significantly reduced when censored observations are imputed. This is true for all three analytes and is shown for tetrachloroethene in Fig. 5. There are two reasons for this. Firstly, percentiles can be transformed back to original scale units without transformation bias. Secondly, since they are more robust towards the

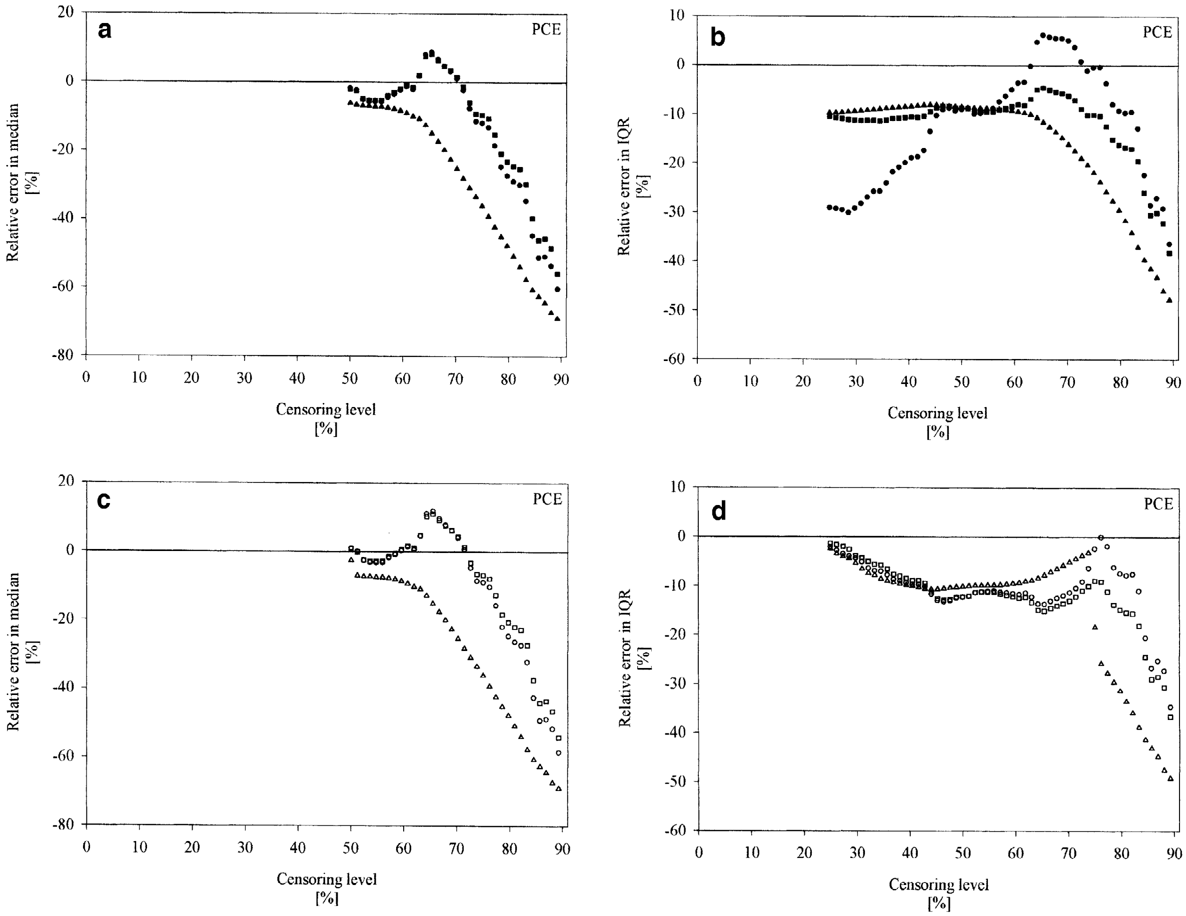

Fig. 5. Comparative performance of statistical methods for estimating the median and interquartile range (IQR) of artificially censored tetrachloroethene data. ● Cohen's ML; ■ bias-corrected restricted ML; ▲ probability-plot regression; ○ robust Cohen's ML; □ robust bias-corrected restricted ML; △ robust probability-plot regression.

presence of outliers, they are less influenced by lack of fit to the lognormal distribution.

4. Conclusion

This paper addressed the problem of estimating summary statistics from censored water quality data sets. Several parametric and robust parametric procedures based on maximum likelihood estimation or probability-plot regression were evaluated in a case study to find the most appropriate method to deal with this issue.

The use of parametric methods requires that the underlying distribution of the data is known. If censored measurements are present, however, assumptions regarding the underlying distribution are hard to check. Although the lognormal distribution usually provides a good description of environmental data sets, the presence of outliers may affect this assumption. Since moments are very sensitive to outliers, lack of fit of outlying measurements to the assumed distribution will result in highly biased estimates of the mean and standard deviation. Moreover, even if the assumptions concerning the underlying distribution are true, moments will be biased due to back-transformation of the estimates to original scale units. Quantiles are more robust to the presence of outliers, and are not subject to these problems.

The implementation of a robust fill-in technique combined with each of the parametric methods significantly reduces bias as far as the moments are concerned. Computation is performed in original scale units, avoiding lack of fit of high values to the lognormal distribution as well as back-transformation of estimates.

To compute summary statistics from the censored data sets observed in the monitoring survey, we suggest the use of robust parametric methods for estimation of moments, while both parametric and robust parametric methods will do for estimation of quantiles. Although results do not point towards one method specifically, the bias-corrected restricted maximum likelihood method and the probability-plot regression method often yielded the least biased results for quantiles, while their robust counterparts gave the best estimates for the mean and standard deviation. Moreover, these estimators are easier to compute than Cohen's maximum likelihood technique. If the recommendations described above are considered, it was shown that summary statistics can be estimated up to 80% of censoring with acceptable bias (5–10%) for the data sets observed in this study.

Acknowledgements

The authors acknowledge financial support by the programme *Sustainable Management of the North Sea* of the Belgian State-Prime Minister's Services-Science Policy Office (contract no. MN/DD/12).

References

[1] Ministerial Declaration of the Third International Conference on the Protection of the North Sea, Den Hague, 1990.

[2] Report of the Marine Chemistry Working Group ICES, Copenhagen, 2000.

[3] Press notice ''Further Protection for the North-East Atlantic'', in: Annual Meeting of the OSPAR Commission, Copenhagen, 2000.

[4] T. Huybrechts, J. Dewulf, O. Moerman, H. Van Langenhove, J. Chromatogr. A 893 (2000) 367.

[5] L.A. Currie, Pure Appl. Chem. 67 (1995) 1699.

[6] R.J. Gilliom, D.R. Helsel, Water Resour. Res. 22 (1986) 135.

[7] M.C. Newman, P.M. Dixon, B.B. Looney, J.E. Pinder, Water Resour. Bull. 25 (1989) 905.

[8] C.N. Haas, P.A. Scheff, Environ. Sci. Technol. 24 (1990) 912.

[9] D.R. Helsel, R.J. Gilliom, Water Resour. Res. 22 (1986) 147.

[10] D.R. Helsel, R.M. Hirsch, in: Statistical Methods in Water Resources, Elsevier Science, Amsterdam, 1992, p. 522, Chapter 13.

[11] S. Kuttatharmmakul, D.L. Massart, D. Coomans, J. Smeyers-Verbeke, Anal. Chim. Acta 441 (2001) 215.

[12] A.H. El-Shaarawi, Water Resour. Res. 25 (1989) 685.

[13] C.C. Travis, M.L. Land, Environ. Sci. Technol. 24 (1990) 961.

[14] H. Ahn, J. Am. Water Resour. Assoc. 34 (1998) 583.

[15] S.W. Hinton, Environ. Sci. Technol. 27 (1993) 2247.

[16] S. Kuttatharmmakul, J. Smeyers-Verbeke, D.L. Massart, D. Coomans, S. Noack, Trends Anal. Chem. 19 (2000) 215.

[17] C.N. Kroll, J.R. Stedinger, Water Resour. Res. 32 (1996) 1005.