INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0127

From monitoring data to regional distributions : A practical methodology applied to water risk assessment.

B. Govaerts, B. Beck, E. Lecoutre, C. Le Bailly and P. Vanden Eeckaut

http://www.stat.ucl.ac.be

From monitoring data to regional distributions : A practical methodology applied to water risk assessment.

B. Govaerts [1,*] B. Beck [2], E. Lecoutre[1],
C. le Bailly[1] and P. Vanden Eeckaut[1]

[1]Université Catholique de Louvain, Institut de Statistique, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.

[2]Eli Lilly, rue Granbonpré 11, B-1348 Mont-Saint-Guibert, Belgium.

August 29, 2001

SUMMARY

During the last decade, the assessment of concentration levels of chemical substances in the environment has become a major issue for defining actual exposure in risk evaluation processes. Currently, different approaches are proposed to achieve this assessment. The first one, widely used, relies on multi-media fate models. On the other hand, monitoring data begin also to be used for this purpose but, up to now, few statistical methods are proposed to validate and summarize them adequately at a local or regional level. A very important characteristic of monitoring data is that many observations stay below the detection limits of the measurement devices and therefore the data are highly censored.

This paper presents a practical methodology to estimate a regional distribution for the concentration of a chemical substance in the surface water of a given region. The estimate of the distribution is obtained from complete or summarized monitoring data collected over time in the region's sampling stations. From this estimate, it is possible to easily derive different statistical summaries as, for examples, the concentration mean, the standard deviation, the percentiles, etc.

Two approaches have been developed and are compared on UK mercury data. A first non-parametric approach derives the regional distribution by aggregating the observed data without taking any assumption on their statistical distribution. A second approach, which may be applied to complete or summarized local data, proceeds in two steps. First, a statistical distribution (e.g. lognormal, gamma,...) for the concentration is fitted to the, possibly censored, data of each sampling station. These local distributions are then aggregated at a regional level and statistics of interest derived.

The paper is closed by a discussion on how to build confidence intervals for the regional summaries and how to fix the weights to combine the distributions from the local to the regional level.

Keywords: Risk assessment, censored data, monitoring data, water quality, maximum likelihood.

*Correspondence to: B. Govaerts, Université Catholique de Louvain, Institut de Statistique, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve (Belgium). E-mail: govaerts@stat.ucl.ac.be Phone: +32-10-47.43.13 Fax: +32-10-47.30.32

1. INTRODUCTION

In order to assess the risk of a chemical substance for a given species and/or on a given ecosystem, two crucial parameters need to be evaluated. The first one, called the Predicted Environmental Concentration (PEC), is used to quantify the actual level of exposure to the substance of interest in the studied system. Whereas, the second value, called the Predicted No-Effect Concentration (PNEC), represents the maximum acceptable non-toxicity level of concentration. The ratio of these two values (PEC divided by PNEC) provides a signal on the risk associated with the given substance. Indeed, if this ratio is above one, actions should be taken in order to reduce potential risk.

For environmental purposes, it is crucial to obtain precise estimations of these two values for the main chemical substances produced by industry. The values of the PNEC for the main industrial substances have been thoroughly studied on the basis of (eco)-toxicological data and their predictions are nowadays considered reliable for many common substances. The main problem then remains in the estimates of the PEC values for the same substances. Two alternatives are available, namely the modelling and the monitoring approaches. The modelling approach is traditionally by far the most used approach in this context but given that the monitoring data are becoming more and more reliable and due to model uncertainties, the monitoring approach can be considered as a serious complement and even an alternative to the modelling approach.

This paper focuses on surface water monitoring data and investigates different alternatives to summarize the data collected over time at the sampling stations of a given region. More precisely, it proposes a methodology to estimate the regional statistical distribution of the concentration from which regional statistics like the mean, standard-deviation or percentiles may be derived. The PEC is often chosen as being a high percentile of this last distribution, usually the 90^{th} percentile.

The paper starts with a clarification of the objectives and with a description of the data set used to illustrate the methodology. Section 3 describes the general methodology, which is based on three different approaches. The first approach, denoted in the following as the "empirical" approach, is presented in Section 4 whereas the two remaining approaches based on parametric statistical distribution are introduced in Section 5 and Section 6. Finally, some additional topics, discussions and comments are given in the last section. All topics are illustrated on mercury monitoring data issued from UK surface water.

2. OBJECTIVES AND DATA DESCRIPTION

2.1. Current practice

As mentioned in the introduction, two approaches are used to evaluate the exposure level of a given chemical substance to human and the environment: the modelling and monitoring approaches. The *modelling approach* is the more traditional one and is used by the authorities to estimate PEC values. Its purpose is mainly to predict the level of concentration of a given pollutant in the different compartments (air, water, soil, sediments or biota) of a geographical region. Three important inputs of these models are the (estimated) emissions of the chemical of interest into water, soil and atmosphere by industries and end users. The second is the (estimated) degradation rates of these chemicals when released in the different compartments. The third are the partition coefficients between the various compartments. The complexity and precision of these models depend of their level of development : local area or global region.

Information like physico-chemical properties of the substances of interest and the properties of the environmental compartments (nature of soil, temperature,UV flux, bacteria population...) is used to predict the evolution of the emitted pollutants in the area of interest. Two well-known examples of such models are EUSES (EUSES, 1996) and GREATER (Feijtel *et al*, 1997). The difference between these two models is mainly that the EUSES model incorporates all types of compartments whereas the GREATER model deals only with rivers networks.

The modelling approach is not without drawbacks. First, the quality of the results strongly depends on the quality of the input data available (emissions, degradation rates, etc.). There is no reason to develop complex models if the input values and estimated parameters are only crude estimates. Secondly, the modelling approach rarely provides full and realistic distributions of concentration and often focuses on a worstcase risk analysis only. Finally, up to now, few effort has been done to validate the predictions of the models by using good quality monitoring data.

The second possible method applicable to evaluate the exposure levels is the monitoring approach. For many years now, most developed countries have organized a monitoring program aimed at periodically measuring the concentration level for a set of pollutants in various compartments (e.g. surface water) at several sampling stations spread over the country. These data are mainly used to check locally whether the concentration level does or does not exceed a fixed alert threshold. They are often gathered together in national databases but are only summarized with basic statistical tools.

One important difficulty in this context is that a lot of chemical concentrations are very low and thus, not detectable by current measurement devices. These data are then reported as "below determination limit" or "left censored" in the statistical literature. Special care must be taken to handle such kind of data in the statistical analysis. Another difficulty comes from the high asymmetric form of the concentration distributions which does not allow to simply apply the statistical methods based on Gaussian distributions. These particularities may explain why, up to now, few statistical methods have been proposed to summarize monitoring data adequately.

2.2. Goal of the paper

This paper proposes a methodology to summarize the concentration levels of chemical substances in the rivers of a given region on the basis of monitoring data. The idea is to estimate a statistical distribution for this regional concentration from which any parameter like the mean, standard deviation, median or other percentiles may be easily derived.

The main motivation of this research is to show how monitoring data can be adequately summarized at a regional level and contribute, as a result, to risk assessment. This perspective is also shared by other organisms like the European Center for Ecotoxicology and Toxicology of Chemical who claims that "monitoring data will play a dominant role in risk assessment if they are judged to be of adequate reliability and representative of the particular environmental compartments to which they are applied" (ECETOC 1999). Note that the methodology has been developed as a result of a research project initiated by the European association of the chlor-alkali industry Euro Chlor and has been applied and validated on several substances of an European monitoring database (COMMPS 98).

2.3. The available database

A lot of data are recorded each year to assess the quality of surface water in the world. Unfortunately, most of these data are currently used at a local or national level only. Few efforts have been made, up to now, to gather them and to unify the measurement methods and database structure.

The European project COMMPS (1998) made the effort to build such a database. Its goal was to establish for the European Commission a list of substances prioritized on the basis of their risk to the aquatic environment and thus indirectly, of their risk to the human health. For this purpose, the countries of the EU have been requested to provide monitoring data for 572 chemical substances. These data have been summarized using basic statistical methods by the COMMPS project team and used to define priority substances to the regulated water framework directive (EC 2000).

The final database collects around 800 000 individual measurements made from 1994 to 1998 at more than 5000 sampling stations spread over 13 countries of the EU. The density of information is not uniform over all EU countries since the monitoring tradition and the policy to make data available considerably varies from one country to another. As a result, more than 50% of the full database originates from UK. Moreover, as more than 70% of the measurements are below the determination limits, the collected data are strongly censored. Note that these determination limits, which depend on both the sampling procedure and the detection limits of used measurement devices, may vary, for the same chemical substance, not only from one station to another but also from day to day in the same station.

For storage requirement and lack of information (some countries did not provide the full information), the COMMPS team decided to build a database with summaries of the individual data. For each substance, the data available were summarized, per sampling station, in 9 statistics : the number of measurements above and below the determination limit, the minimum and maximum determination limit, the minimum, mean, median, maximum and standard deviation of the observations above the determination limit.

Deliberately, the proposed methodology is based on this peculiar structure of the database. For practical consideration, it seems indeed more realistic to consider that in the future, it will be easier to obtain summarized data than individual data. Nevertheless, for validation purposes, a methodology for dealing with individual data was also developed.

Although the method proposed below has been applied to all chlorine and mercury related substances available in the COMMPS database it will be illustrated here only on the mercury data collected in UK rivers. The full results of the analysis are now used by the Euro Chlor federation in Europe.

2.4. First analysis of UK mercury data

This paper illustrates the methods on total mercury data collected in UK surface water from 1994 to 1996. 16861 individual data were collected over 489 sampling stations corresponding to about 35 measurements per sampling station. All data are given in μg/l. In this data set, 65% of the measurements are censored with a very wide range of determination limits. Table 1 shows the list of the most frequent determination limits with the corresponding relative frequency in the database.

It is not easy to provide a descriptive summary of a data set containing censored values and the authors have dedicated a part of their research to develop a methodology to perform such analysis when only summarized data are available. It aims also at checking if outliers and non-coherent information are present in the data. Details may be found in Govaerts and Vanden Eeckaut (1999).

As an example, to get simply an idea of the mean concentration over all sampling stations, several approaches may be considered: one may discard censored data and compute the arithmetic mean on the remaining data. Another possibility is to replace the censored data by an arbitrary value like 0, or by the half of the value of

Detection limit frequency		
DL (μg/l)	n	%
0.008	1382	12
0.01	2566	23
0.02	4872	44
0.03	1090	10
0.1	1119	10
Others	19	< 1

Table 1: The most frequent determination limits (DL) for UK mercury data (μg/l)

the determination limit and compute the usual arithmetic mean over the observed and imputed data.

The first approach (used for many substances in the COMMPS project) will usually lead to an overestimation of the true mean. Although the second approach uses more information, a replacement by 0 will clearly underestimate the true mean and, on the contrary, a replacement by DL would overestimate it. Using half of the determination limit is probably a good compromise, which is commonly used but not without criticisms (El-Shaarawi and Esterby 1992).

For the mercury data these means are given in Table 2. A same weight has been given to each observation in the mean computation.

Computation method	Estimated Mean (μg/l)
Discard censored data	0.186
Replace censored data by 0	0.064
Replace censored data by DL/2	0.073
Replace censored data by DL	0.081

Table 2: Estimation of the mean concentration of UK mercury data for different imputation of censored observations

In this table, no mean is the "true" arithmetic mean but it indicates at least that this value is certainly in the interval [0.064,0.081]. Figure 1 represents a histogram of all data. Censored data have been replaced by half of the corresponding determination limit.

Note that observed concentrations are in most cases very low and often close to determination limits. Some outliers around 35 μg/l are observed in the data set. They will be kept in the analysis to illustrate the method's sensitivity to such observations. Moreover, 34 stations with less than 3 observations have been removed from the analysis because they can not be considered as informative enough to be included.

2.5. Notations

In this paper, the following notation is used :

R: the region of interest.

r: the number of sampling stations in the region R.

S_i: the i^{th} station of region R ($1 \leq i \leq r$).

X_i: the random variable representing the concentration of the chemical of interest in station i.

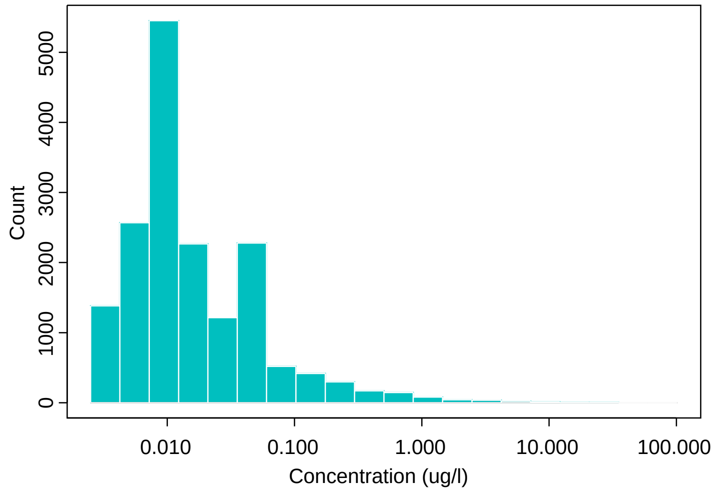

Figure 1: Histogram of UK mercury data in log scale. Censored data are replaced by DL/2

m_i and v_i^2: the theoretical mean and variance of X_i.

m_R , v_R^2 and x_p^R: the theoretical regional mean, variance and pth percentile of the concentration.

n_i: the number of measurements made in station S_i over the sampling period.

x_{ij}: the j^{th} measured concentration in station S_i $(1 \leq j \leq n_i)$.

δ_{ij}: the determination limit attached to x_{ij}. If $x_{ij} \leq \delta_{ij}$, x_{ij} is censored and δ_{ij} is the only information available.

k_i: the number of censored data in station S_i.

C_i^c: the list of indices of measurements of S_i which are censored : $C_i^c = \{j \mid 1 \leq j \leq n_i \quad \text{and} \quad x_{ij} \leq \delta_{ij}\}$.

C_i: the list of indices of measurements of S_i which are not censored.

$\min_i$, med_i, $\max_i$, $\bar{X}_i$, s_i: the minimum, median, maximum, mean and standard deviation of non-censored data in S_i.

$\delta_i^{\textbf{med}}$: the median of the determination limits attached to the censored data in S_i.

N: the total number of measurements in region R : $N = \sum_{i=1}^{r} n_i$.

w_i: the weight allocated to station S_i in data aggregation.

3. GENERAL DESCRIPTION OF THE METHODOLOGY

In this section, we introduce three different approaches proposed to estimate a statistical distribution for the concentration of a substance in the rivers of a given region. Two of these three approaches are based on complete data whereas the last one is based on summary data.

The first approach (denoted as "empirical") is fundamentally nonparametric as the regional distribution is derived without taking any assumption on the local statistical distribution of the data. Data from each site are first gathered together using adequate weights. Then, the regional density and distribution function are estimated numerically using classical nonparametric tools like, for example, the kernel method. All statistics of interest (mean, standard deviation and percentiles) may then be derived from these functions. In this empirical approach, censored data are handled by simple imputation. They are replaced by arbitrary values (e.g. half the determination limit) and considered then as observed. The empirical method is illustrated in Figure 2 and discussed in Section 4.

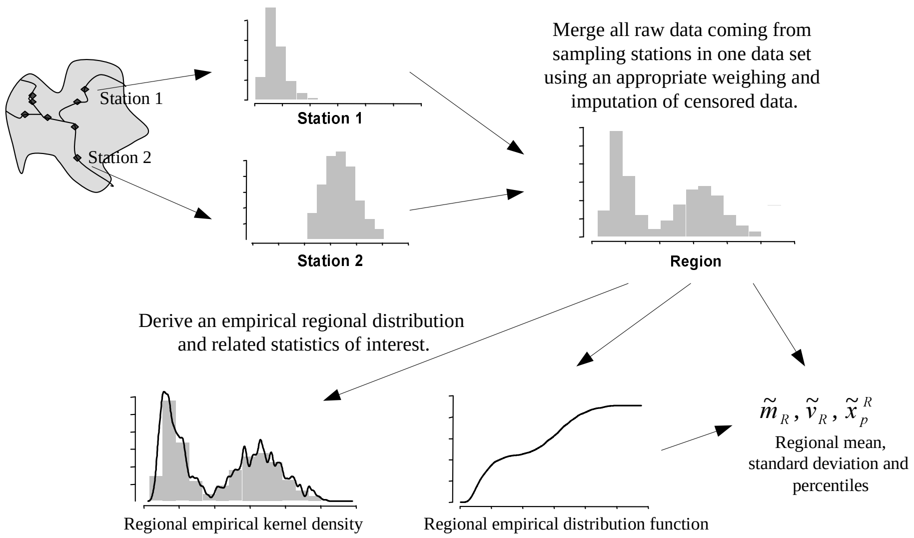

Figure 2: Empirical method to derive a regional distribution

The two other approaches (one for complete data and another for summaries) are parametric since they now suppose that the form of the statistical distribution associated to the data of each sampling station is known. They proceed in two steps.

First, a given statistical distribution (e.g. lognormal, gamma...) is fitted by maximum likelihood to the data of each station. The maximum likelihood method is, in this context, an attractive estimation method since it can be used for any class of parametric distribution, can handle multiple censored data, can be adapted to summary data (see below) and gives more efficient estimations than other methods (Gilliom and Helsel 1986 and Govaerts *et al* 2000).

Then, the regional density is simply defined as a weighted sum of local concentration densities. From this regional density, the regional distribution function, the regional moments and percentiles may be easily derived. The two steps of the parametric method are illustrated in Figure 3 and are discussed in details in sections 5 and 6.

Note that the two solutions based on complete data are interesting by themselves but will also serve as benchmark for the regional distribution obtained from summary data (see sections 6.3 and 6.4).

4. EMPIRICAL REGIONAL DISTRIBUTION

The more transparent way to summarize monitoring data over a region consists in constructing an empirical regional distribution without making any parametric

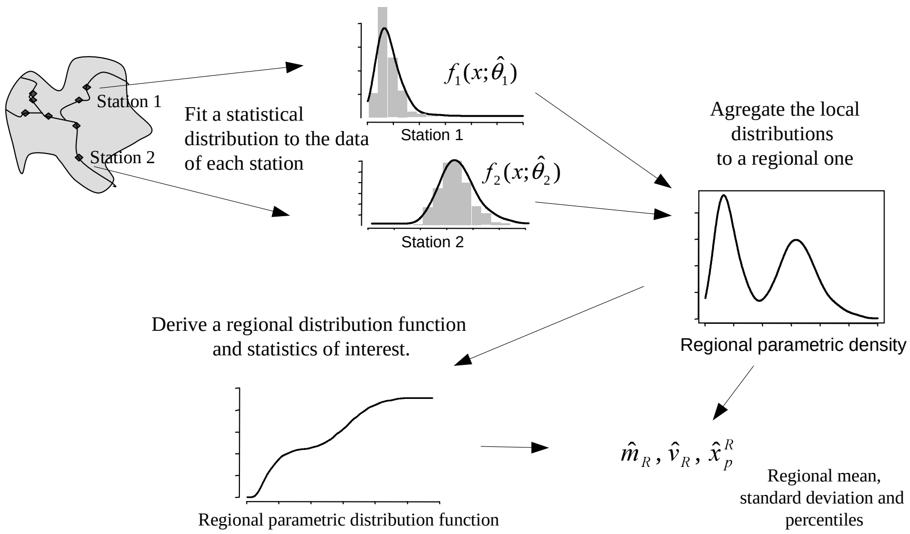

Figure 3: Parametric method to derive a regional distribution

assumption on their underlying distribution.

4.1. Imputation and weighting

Since the data over a region are grouped by sampling station and may contain censored data, two arbitrary choices have been made in the construction of a regional empirical distribution.

First, the data below the determination limit have been replaced by artificial data and treated later as if they were observed. Three imputation strategies were adopted and compared :

DL case Each censored observation x_{ij} is replaced by the corresponding determination limit δ_{ij}. This case leads clearly to an overestimation of the true concentrations but is a conservative choice used for comparison purpose.

DL/2 case Each censored observation is replaced by half of the corresponding determination limit δ_{ij}. This approximation is often used in the literature (El-Shaarawi and Esterby 1992).

UNIF case Each censored observation is replaced with a random number drawn from an uniform distribution from zero to the determination limit δ_{ij}.

Second, the number of observations n_i available for each sampling station may vary from one station to another but, in the aggregation, the weight of each station should be identical. Consequently, when a location has more observations than another, the weight of its observations should be reduced accordingly. Formally, for any location S_i with n_i observations a weight $p_i = 1/n_i$ will be attached to a given observation x_{ij}.

Moreover, in the aggregation, the methodology gives the possibility to allocate an additional weight w_i to each station ($0 \leq w_i \leq 1$ and $\sum w_i = 1$). It allows to express which portion of the rivernetwork each station is supposed to represent and may be based on hydrological and geographical information as discussed in Section 8.3. When such information is not available, the only possible choice is $w_i = 1/r$, which indicates that the sampling stations are well spread over the rivers of the region.

4.2. Construction of an empirical regional density and distribution function

An empirical regional density may be derived by the kernel method which is a smooth version of the histogram. The kernel method consists in attaching to each observation x_i of a sample $x_1, x_2, \ldots, x_n$, a density $K(.)$ (the *kernel*) centered on x_i and spread according to a given *bandwidth* h. The kernel density is then defined as:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}).$$

If a standard normal kernel is chosen, the empirical regional density takes the following form :

$$\tilde{f}_R(x) = \sum_{i=1}^{r} w_i p_i \sum_{j=1}^{n_i} f_N(x; x_{ij}, h^2)$$

where

$$f_N(x; x_{ij}, h^2) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{(x-x_{ij})^2}{2h^2}}.$$

Since concentration data are usually highly left skewed, it is recommended to derive this density on the log of the data. The parameter h will be chosen accordingly to the smoothness wanted for the density. Methods are available to derive optimal h (see e.g. Silverman 1986).

Additionally, an empirical regional (cumulative) distribution function $\tilde{F}_R(x)$ will be defined as the weighted count of the number of observations below a concentration level x.

$$\tilde{F}_R(x) = \sum_{i=1}^{r} w_i p_i \sum_{j=1}^{n_i} I\{x_{ij} \leq x\}$$

where

$$I\{x_{ij} \leq x\} = \begin{cases} 1 & \text{if} \quad x_{ij} \leq x \\ 0 & \text{if} \quad x_{ij} > x. \end{cases}$$

Let us recall that in these two functions, censored x_{ij} are replaced by δ_{ij}, $\delta_{ij}/2$ or a uniform random number between 0 and δ_{ij}.

4.3. Deriving empirical regional statistics

The empirical regional mean $\tilde{m}_R$ and variance $\tilde{v}_R^2$ are defined respectively as :

$$\tilde{m}_R = \sum_{i=1}^{r} w_i p_i \sum_{j=1}^{n_i} x_{ij}$$

$$\tilde{v}_R^2 = \sum_{i=1}^{r} w_i p_i \sum_{j=1}^{n_i} (x_{ij} - \tilde{m}_R)^2$$

where as in the previous sections, censored x_{ij} are imputed.

The percentiles can be derived by inverting the empirical distribution function :

$$\tilde{x}_p = \min\{x \in \mathbf{R}^+ | \tilde{F}_R(x) \geq p\}.$$

Note that for high probabilities p (p bigger than the censored level), the percentile $\tilde{x}_p$ is usually not affected by censored data. Indeed, when a percentile is above all determination limits, its estimation is not any more influenced by the imputation method.

4.4. Illustration on UK mercury data

Empirical densities and corresponding distribution functions for UK mercury data are given in Figures 4 and 5 for the three imputation methods DL, DL/2 and UNIF. Table 3 gives the empirical statistics.

Imputation method	Mean m_R	Median $x^R_{0.5}$	$x^R_{0.75}$	$x^R_{0.9}$	$x^R_{0.95}$
DL case	0.0734	0.0200	0.040	0.100	0.171
DL/2 case	0.0654	0.0130	0.040	0.080	0.170
Uniform case	0.0654	0.0157	0.033	0.088	0.170

Table 3: Empirical regional moments and percentiles for UK mercury data using 3 possible imputation methods for censored data (in μg/l).

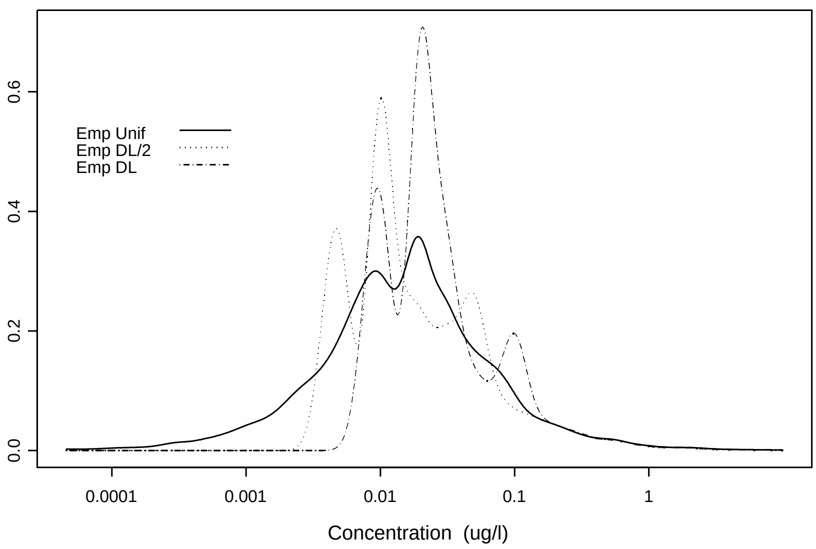

Figure 4: Empirical regional densities for UK mercury data using 3 possible imputation methods for censored data

The following remarks come out of these results:

- The choice of the imputation method has a noticeable effect on the resulting distribution and empirical mean. The three results are clearly differentiated.
- Peaks in DL and DL/2 densities and stairs on corresponding distribution functions arise from very frequent determination limits. These are less present when uniform imputation is applied.
- Small peaks and stairs in uniform imputation curves are explained by the fact that many measured concentrations have been rounded to two decimals. Rounded numbers (0.01, 0.02,...) are then appearing very frequently. This has very little impact on estimated statistics.
- Logically, the DL empirical distribution function is situated to the right of

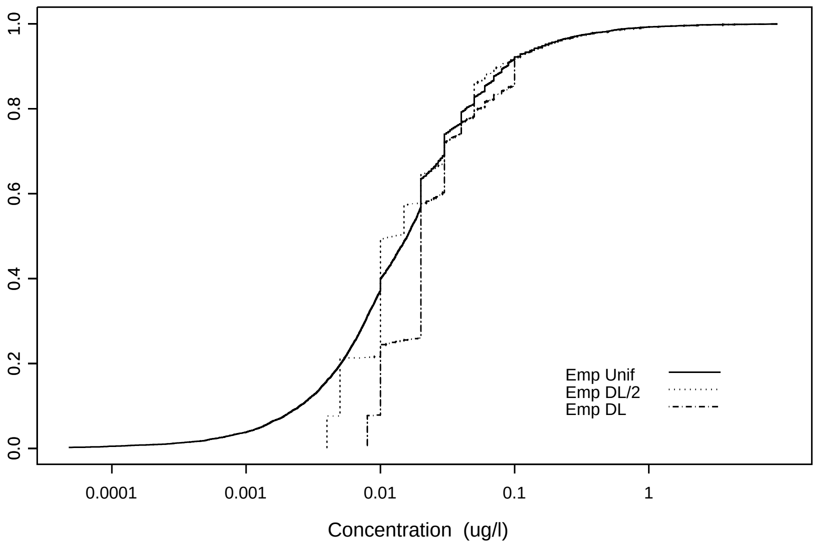

Figure 5: Empirical regional (cumulative) distribution functions for UK mercury data using 3 possible imputation methods for censored data

the DL/2 one and gives, in consequence, higher estimations of concentration percentiles.

- The UNIF empirical distribution is on the left of the DL/2 one for low concentrations and on the right when arriving to percentile 0.4-0.5. This is also logical because a uniform distribution in [0,DL] has around half of the data below DL/2 and half above DL/2.
- For high concentrations, the three empirical densities, distributions and percentiles coincide because they are not any more affected by data below determination limits.
- Due to the high asymmetry of the distribution of the data, the means are very close to the 90^{th} percentile. This is often encountered in environmental data.
- Note finally that the results of Tables 2 and 3 are different because no weighing of individual observations was done in Table 2.

5. ESTIMATION OF A LOCAL DISTRIBUTION

The two step procedure proposed to estimate a regional distribution consists in fitting first a (parametric) statistical distribution $f_i(x;\theta_i)$ to the data of each sampling station S_i and then aggregating these local distributions to a regional one. Two problems are encountered in the estimation of a local distribution. First, the choice of an appropriate family of distributions and secondly, the choice of an estimation method to fit the distribution to the censored data.

In the water quality literature, the lognormal distribution is widely accepted as a good candidate to fit substance concentrations in water. It is well adapted to the left skewed shape of the distribution of such data, it is easy to fit (when data are not censored) and simulation results show that it is quite robust to distribution misspecification (Gilliom and Helsel 1986). The authors confirmed, on several data sets, the good adequacy of the lognormal distribution when fitting concentrations of chemicals in water. Nevertheless, the methodology discussed in this paper applies to other possible distributions too and will therefore be illustrated on both the lognormal and the gamma distributions.

Let us recall that the density function of a lognormal(μ,σ^2) is defined by :

$$f_{LN}(x;\mu,\sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right)$$

and of the gamma(α,β) :

$$f_G(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}.$$

The mean and variance associated with these two distributions are : for the lognormal

$$E(X) = m = e^{\mu+\frac{\sigma^2}{2}} \text{ and } V(X) = v^2 = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

for the gamma

$$E(X) = m = \frac{\alpha}{\beta} \text{ and } V(X) = v^2 = \frac{\alpha}{\beta^2}.$$

Figure 6 presents the shape of these two distributions for different combinations of their parameters.

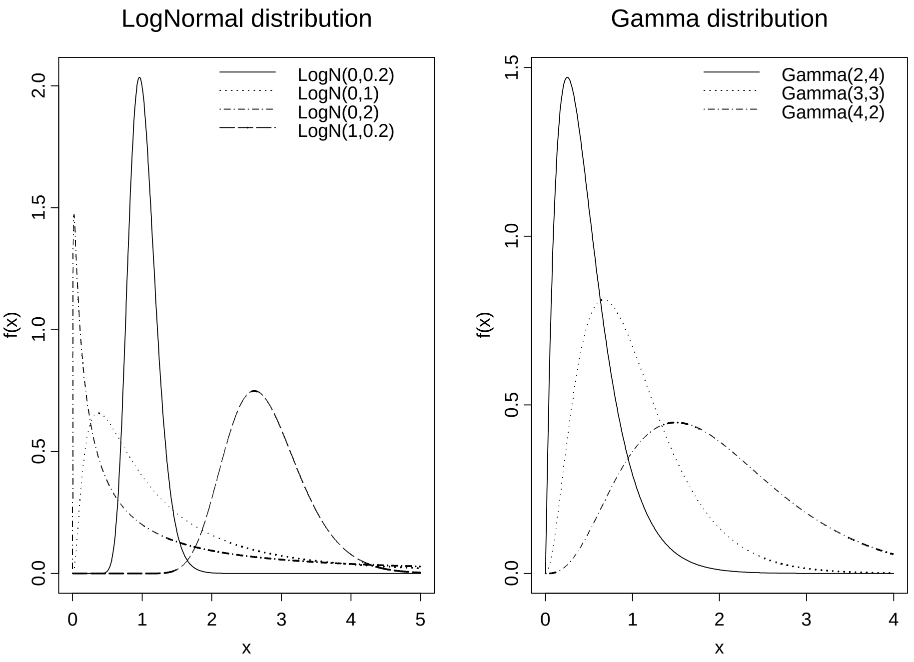

Figure 6: Examples of lognormal and gamma densities

5.1. Available methods

A wide literature is devoted to censored data and many papers in environmental science and water quality journals discuss the problem of summarizing and fitting statistical distributions to left censored data (see review papers like Gilliom and Helsel (1986), El-Shaarawi and Esterby (1992) or Akritas *et all* (1994)). Review papers usually classify the available methods in three categories : simple substitution, parametric and robust parametric methods.

Simple substitution methods replace censored observations by a constant (e.g. half of the determination limit) and combine substituted values with observed values to calculate the summary statistics of interest.

Parametric methods fit a given statistical distribution to the (censored) data and calculate the summary statistics from the estimated distribution. The maximum likelihood method for normal or lognormal censored samples is the most popular method in this category (see Cohen (1951), Gleit (1985) or Wolynetz (1979)).

Robust parametric methods construct values for the censored data by fitting a parametric distribution to the uncensored data and extrapolating it to fill in the censored observations. Like for substitution methods, the completed sample is used to calculate the summary statistics of interest. For example, the lognormal probability regression is a well known method in this class (Gilliom and Helsel, 1986).

Simulation studies show that the best methods to summarize and fit a distribution to left censored data are the maximum likelihood (ML) method and the log probability regression (LPR) (Gilliom and Helsel (1986) and Govaerts *et al* (2000)). When data are lognormally distributed, ML gives more efficient estimations of the distribution parameters than LPR but LPR seems to be more robust to departure from lognormality than ML. The ML method is preferred in this paper because it is applicable to any class of distribution and handles easily multiple determination limits.

5.2. The maximum likelihood method for censored data

Let $\chi_i' = (x_{i1}, x_{i2}, \ldots, x_{in_i})$ be a sample of measured concentrations of a given substance in station S_i. In this sample, a subset C_i^C of k_i observations are censored. The determination limit δ_{ij} is the only available information. If $f_i(x;\theta_i)$ is the density function attached to these data, the likelihood function for censored and uncensored data may be written as (Cohen 1951) :

$$L(\chi_i;\theta_i) = \prod_{j\in C_i} f_i(x_{ij};\theta_i) \prod_{j\in C_i^c} F_i(\delta_{ij};\theta_i)$$

where $F_i(x;\theta_i)$ is the cumulative distribution function associated to the density $f_i(x;\theta_i)$:

$$F_i(\delta_{ij};\theta_i) = \int_0^{\delta_{ij}} f_i(x;\theta_i)dx.$$

The likelihood takes the following form when the data are lognormally distributed:

$$L(\chi_i;\mu_i,\sigma_i) = \prod_{j\in C_i} \frac{1}{x_{ij}\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(ln(x_{ij})-\mu_i)^2}{2\sigma_i^2}\right) \prod_{j\in C_i^c} \int_0^{\delta_{ij}} f_{LN}(x;\mu_i,\sigma_i^2)dx.$$

Several methods may be used to derive maximum likelihood estimators for the distribution parameters. The methods are, in most cases, numerical iterative algorithms since, for usual classes of distributions, no analytical solution exists for this maximization problem when data are censored and multiple determination limits may be present.

For the lognormal case, Cohen (1951) and Wolynetz (1979) have derived two variants of a system of two (nonlinear) equations to be solved to get the parameter's estimations. Cohen (1951) provides also numerical tables to find a solution in one stage. Le Bailly (2000) shows that the Wolynetz equations give much more stable solutions when solved numerically. These two methods have the important drawback that they do not support multiple detection limits and concern only the lognormal distribution. A general and very robust solution consists in performing simply a grid search of the optimal solution in the parameter's domain. Such an algorithm is made efficient by using two or three embedded grids to isolate the optimum and gives very comparable results to the Wolynetz method in the lognormal case.

5.3. The maximum likelihood method for summary data

When dealing only with summary data in each sampling station, a likelihood function may be defined in the same spirit as the expression used for censored data. In the context of this paper, since the minimum, mean, median, maximum of non-censored data and the median of determination limits are available, the following form is proposed for the likelihood:

$$\begin{aligned} L(\chi_i;\theta_i) &= F(\delta_i^{\mathrm{med}};\theta_i)^{k_i} f(\mathrm{min}_i;\theta_i)(F(\mathrm{med}_i;\theta_i)-F(\mathrm{min}_i;\theta_i))^{\frac{n_i-k_i-3}{2}} \\ & \quad f(\mathrm{med}_i;\theta_i)(F(\mathrm{max}_i;\theta_i)-F(\mathrm{med}_i;\theta_i))^{\frac{n_i-k_i-3}{2}} f(\mathrm{max}_i;\theta_i). \end{aligned} \tag{1}$$

This formula expresses the fact that three values are observed (min,max,med) and all others are known to be in given intervals : k_i censored values are below or closely above δ_i^{med}, $\frac{n_i-k_i-3}{2}$ between the observed minimum and median and $\frac{n_i-k_i-3}{2}$ between the median and the maximum. This is illustrated in Figure 7. Numerical solutions of the maximization problem may be found by grid search.

The choice of δ_i^{med} as a summary of the detection limits attached to the censored data of station S_i is a way to approach as good as possible the corresponding expression in the likelihood for complete data :

$$F(\delta_i^{\mathrm{med}};\theta_i)^{k_i} \cong \prod_{j\in C_i^c} F_i(\delta_{ij};\theta_i).$$

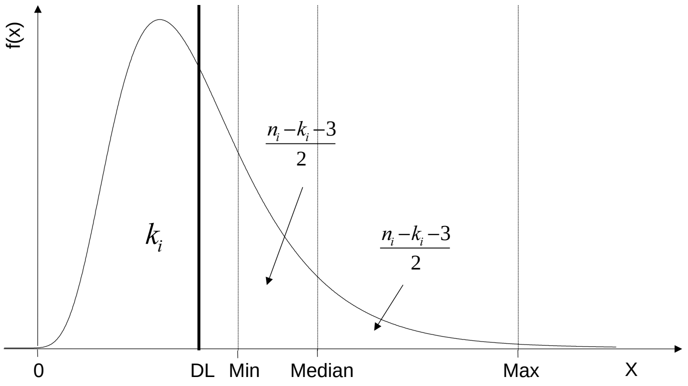

Figure 7: Structure of the likelihood function based summaries

Lebailly (2000) and Govaerts *et al* (2000) have made a comparison of the quality of the estimations obtained from complete or summary data. This comparison is very encouraging since it shows that the summary method gives estimations nearly as efficient than the complete method.

5.4. Illustration on UK mercury data

A data sample of mercury concentrations in a sampling station of UK is used here to illustrate and compare the different approaches discussed above to fit a local distribution to censored data. This sample consists of 153 observations among which 62 are censored with three determination limits 0.02, 0.03 and 0.1 μg/l. Lognormal and gamma distributions have been fitted to the complete and summary data. The estimated parameters are given in Table 4, and Figure 8 shows the estimated distributions over a histogram of the data in real values (left) and in log (right). In the histogram, the censored data have been replaced by half of their determination limit.

		μ	σ
Lognormal	All data	-3.492	1.211
	Summaries	-3.534	1.338
		α	β
Gamma	All data	0.902	13.801
	Summaries	0.639	10.118

Table 4: Estimated parameters of four local distributions fitted to the data of a single sampling station in the UK

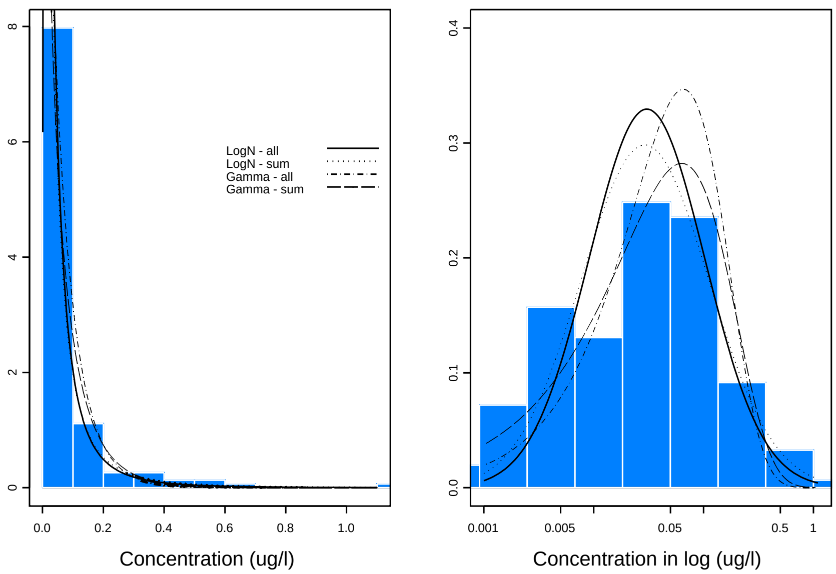

Figure 8: Comparison of four local distributions fitted to the data of a single sampling station in UK.

This example shows that the distributions estimated with complete and summary data are very similar and shows that the differences between the lognormal and gamma approaches are well marked in the shapes of the distribution's tails. Also, parameters of the lognormal based on all data and summary data are closer to each other than gamma ones. This has been observed on several other data sets.

6. AGGREGATION OF LOCAL DISTRIBUTIONS TO A REGIONAL ONE

6.1. Construction of a parametric regional density and distribution

When a (parametric) local density $f_i(x;\hat{\theta}_i)$ has been estimated for each sampling station S_i on complete or summary data, a regional parametric density can be defined as :

$$\hat{f}_R(x) = \sum_{i=1}^{r} w_i f_i(x;\hat{\theta}_i)$$

where w_i is a weight attributed to station S_i taking into account its importance in the region.

The corresponding distribution function may be derived as :

$$\hat{F}_R = \sum_{i=1}^{r} w_i F_i(x;\hat{\theta}_i).$$

Since this distribution has rarely an analytical form, the best way to handle it consists in calculating the value of $\hat{F}_R(x)$ on a grid defined over the concentration domain. For a good resolution, a grid equally spaced in log is a reasonable choice.

Note that the classes of local densities do not need to be the same for all stations. In principle, one may select for each station the class of distribution that fits better the data but in practice, it seems reasonable to choose a common class for each main river of the region of interest.

6.2. Deriving parametric regional statistics

Estimators for regional statistics can be derived analytically or numerically from the regional distribution defined above. An estimator for the regional mean m_R is given by :

$$\hat{m}_R = \int_0^{\infty} x\hat{f}_R(x)dx = \sum_{i=1}^{r} w_i \hat{m}_i$$

where $\hat{m}_i$ is the estimated mean concentration in station i. When a lognormal distribution is fitted locally by maximum likelihood, $\hat{m}_i$ in given by :

$$\hat{m}_i = e^{\hat{\mu}_i + \frac{\hat{\sigma}_i^2}{2}}.$$

The regional variance σ_R^2 can be estimated as

$$\hat{v}_R^2 = \int_0^{\infty} (x-m)^2 \hat{f}_R(x)dx = \sum_{i=1}^{r} w_i \hat{v}_i^2 + \sum_{i=1}^{r} w_i(\hat{m}_i - \hat{m}_R)^2.$$

Regional percentiles (median, $x_{0.9}^R$...) can only be estimated numerically from the regional (cumulative) distribution function $\hat{F}_R(x)$ and are defined as follow :

$$\hat{x}_p^R = \min\{x \in \mathbf{R}^+ | \hat{F}_R(x) \geq p\}.$$

6.3. Illustration on UK mercury data

Aggregated distributions are presented here for UK mercury data described in Section 2.3. Four regional distributions have been estimated from the 455 local distributions. The first is based on local lognormal distributions fitted on complete data (LogN-all), the second on lognormal with summary data (LogN-sum) and the two others on local gamma distribution fitted on complete and summary data (Gamma-all and Gamma-sum).

Figures 9 and 10 show the estimated parameters of the 455 local distributions in these four situations. In each graphic, the pairs of parameters obtained from complete (All) or summarized data are linked together. One can see that the lognormal local parameters obtained from complete and summarized data are in most cases very similar. This is less true for the gamma ones were parameters based on summaries are often more extreme than those based on complete data. This will result in more instability in the regional results.

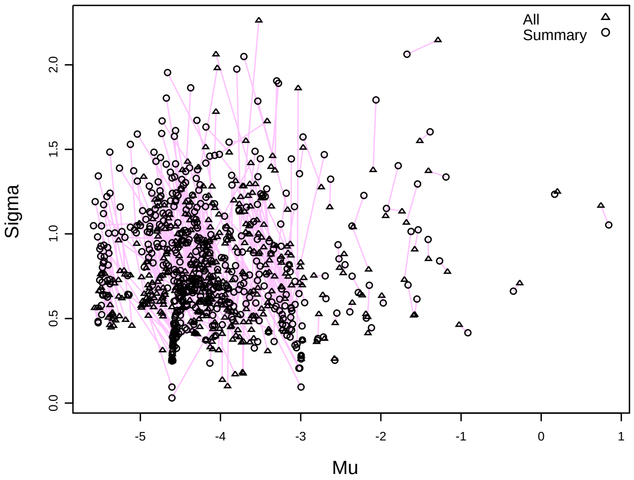

Figure 9: Local parameters of the lognormal distribution estimated on complete and summary UK mercury data in the 455 sampling stations

Figure 11 compares the four densities and Figure 12 the corresponding distribution functions. Table 5 gives the regional statistics.

Estimation	method	Mean m_R	Median $x^R_{0.5}$	$x^R_{0.75}$	$x^R_{0.9}$	$x^R_{0.95}$
Lognormal	All data	0.0641	0.0161	0.0338	0.0802	0.1524
	Summaries	0.0639	0.0150	0.0363	0.0810	0.1546
Gamma	All data	0.0585	0.0167	0.0366	0.0781	0.1365
	Summaries	0.0585	0.0152	0.0358	0.0810	0.1651

Table 5: Parametric regional moments and percentiles for UK mercury data using 4 possible methods to estimate local distributions (μg/l)

The comparison of the distributions and statistics based on complete and summary data show little loss in working with summaries. This is specially true when lognormal local distributions are used. This may have in practice a big impact on data storage. Second, lognormal and gamma based results are also close to each other. A difference may be seen in the left tails of the distributions but it is inflated by the log scale of the graphic and has little impact on estimated regional moments and percentiles.

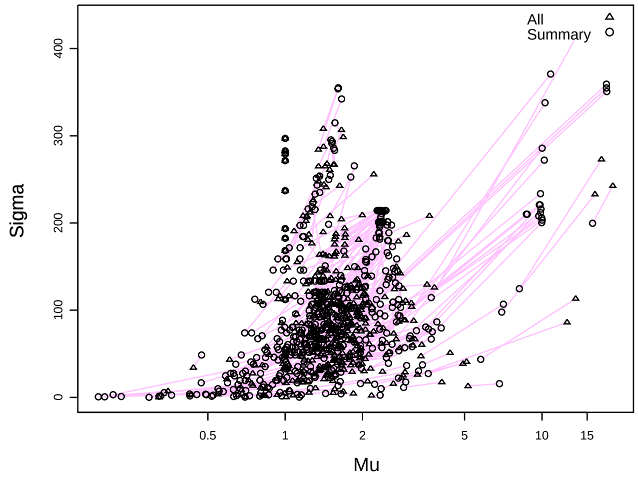

Figure 10: Local parameters of the gamma distribution estimated on complete and summary UK mercury data in the 455 sampling stations

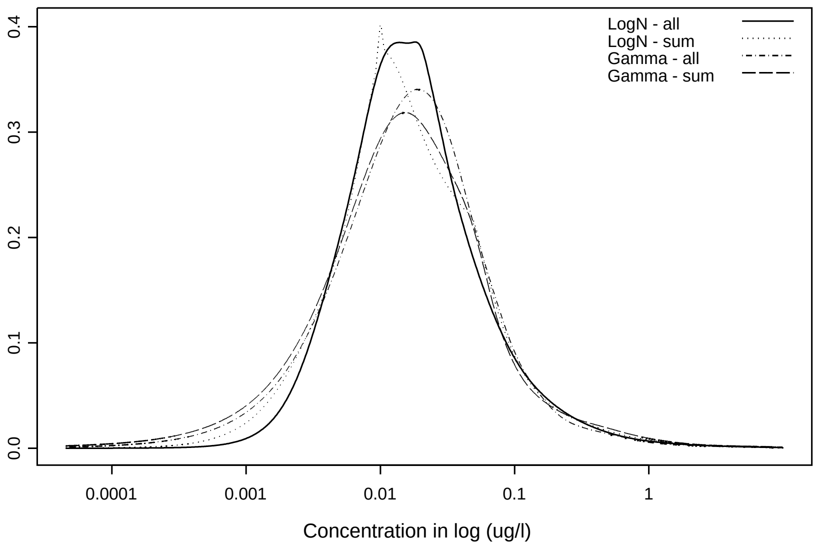

Figure 11: Parametric regional densities for UK mercury data using 4 possible methods to estimate local distributions

7. COMPARISON OF THE DIFFERENT APPROACHES ON MERCURY DATA

The empirical and parametric methods presented in sections 4, 5 and 6 to build the regional concentration distribution are compared here on the UK mercury data.

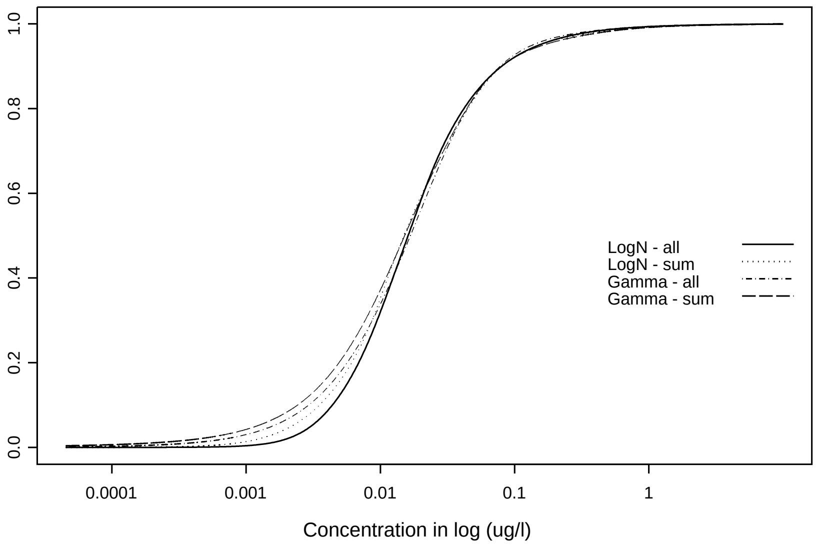

Figure 12: Parametric regional distribution functions for UK mercury data using 4 possible methods to estimate local distributions

Since the previous section shows that the distribution obtained from the complete and summary data are sensibly the same, the comparison is limited in this section to the parametric regional distributions based on complete data. For empirical data, substitution by $DL/2$ and uniform imputation are presented.

Figure 13 compares the densities obtained in the 4 situations, Figure 14 the distribution functions and Table 6 the regional statistics.

Estimation	method	Mean m_R	Median $x^R_{0.5}$	$x^R_{0.75}$	$x^R_{0.9}$	$x^R_{0.95}$
Empirical	DL/2 case	0.0544	0.0130	0.0400	0.0800	0.1700
	Uniform case	0.0654	0.0157	0.0400	0.0877	0.1700
Lognormal	All data	0.0641	0.0161	0.0338	0.0802	0.1524
Gamma	All data	0.0585	0.0167	0.0366	0.0781	0.1365

Table 6: Comparison of empirical and parametric regional moments and percentiles for UK mercury data (μg/l)

These results show that both parametric distribution functions are good approximations of the empirical ones. They both tend to slightly underestimate the high percentiles and, globally, the results based on lognormal local distributions give better results than those based on gamma ones. More generally, other data sets have shown that the distributions based on gamma local distributions are less stable than those based on lognormal.

In conclusion, the parametric approach seems to provide a satisfactory approach to derive a regional distribution and related statistics. Its advantage over the empirical approach is that it does not rely on an imputation method and allows for more efficient data storage since the data of each sampling station may be summarized by two estimated parameters.

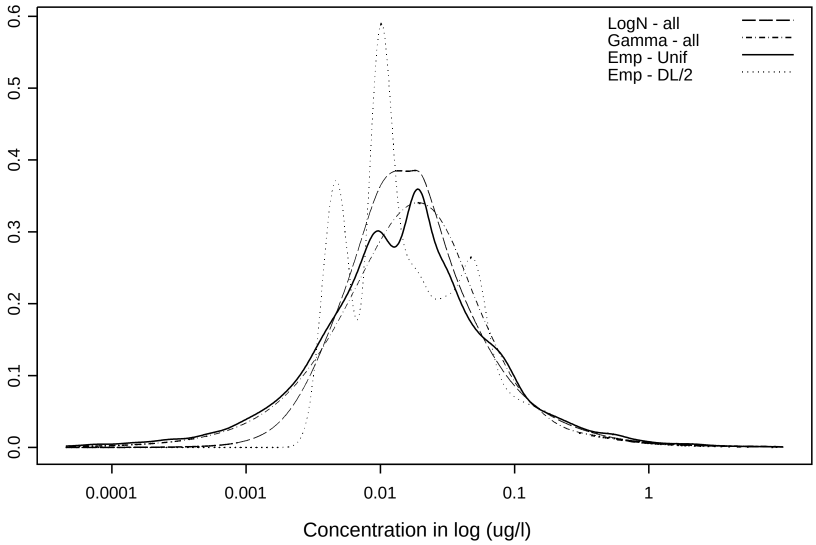

Figure 13: Comparison of empirical and parametric regional densities for UK mercury data

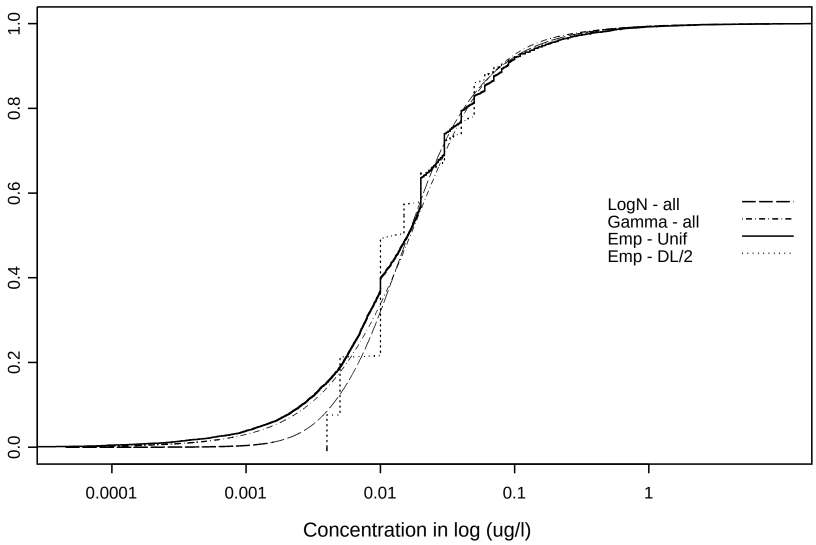

Figure 14: Comparison of empirical and parametric regional distribution functions for UK mercury data

8. FURTHER DEVELOPMENTS

This section discusses some open questions linked to the methodology presented in this paper. They are related firstly to the estimation of local parametric distributions, secondly, to the derivation of confidence intervals for the regional parameters of interest and finally, to the choice of the weights in the aggregation of the local information to regional results.

8.1. Improving the method to estimate local distributions

The parametric approach to derive a regional distribution consists in aggregating the local distributions fitted to the data of each sampling station. The quality and reliability of the regional results will then crucially depend on the capacity of the fitted local distributions to mimic the "real" distributions of the concentration. In order to improve the approach discussed in Section 5, three questions would merit further investigations.

Firstly, the methodology allows the choice of a specific class of local distribution $f_i(x;\theta_i)$ for each location if it seems necessary. In practice, goodness of fit tests should systematically be used to find the most appropriate family of distributions for the stations of each main river of interest (lognormal, Weibull, gamma,...). This is more difficult in the presence of censored data with multiple censored limits. Methods discussed in Akritas (1988) or D'Agostina and Stephens (1986) should be tested and compared on water quality data.

Secondly, more work is necessary to test and improve estimation methods based on summary data. Le Bailly (2000) has already compared by means of simulations the maximum likelihood estimators based on complete and summary data and has shown that the summary method gives estimations nearly as efficient as the complete method. Nevertheless, the robustness of the maximum as a summary statistic may seriously be questioned when one may suspect the presence of outliers, and other types of order statistics like quartiles would merit to be tested as potential summaries. Moreover, the literature shows that the log probability regression (LPR) estimation method is more robust to departure of lognormality (Gilliom and Helsel 1986). One may then attempt to adapt the LPR to summary data and check if efficiency and robustness is still observed.

Finally, it is common knowledge that the maximum likelihood estimators for censored data are biased in small samples when censored level is high. Currently, the methodology does not propose any bias correction because the authors think that this has little impact on the final result since they showed through simulations in the lognormal case that this bias is small compared to the variability of the estimators. Nevertheless, their simulations did not cover all possible situations and further investigation are needed.

8.2. Building confidence intervals on regional parameters

A natural complement to the methodology discussed above consists in providing confidence intervals with the estimated regional parameters of interest in order to reflect their reliability. The solution to this problem is not straightforward and this section will only initiate the discussion.

The first difficulty comes from the fact that a large number of uncertainty sources may affect the estimated statistics. The sampling methodology, the representativeness of the station network, the errors and detection limits of measurement devices and the oversimplification of statistical hypothesis are examples of factors which may have a big impact on final results. The usual statistical confidence intervals suppose that data are reliable, representative for the population of interest and that the postulated statistical hypotheses are verified. They focus only on sampling incertitude due to the fact that only a sample of observations are available from the whole population.

This last point has been investigated by Lecoutre (2001) and is discussed in some details below.

The second difficulty in the derivation of such "classical" confidence intervals comes from the censored data and the fact that regional statistics estimators are either non-parametric, or combinations of parametric estimators obtained numerically. In such situations, no classical small sample method may be used to derive confidence intervals and asymptotic formulae or the bootstrap method are the only possible approach. With such methods there is *a priori* no certitude that resulting intervals cover the parameter of interest with the adopted confidence level.

Lecoutre (2001) has proposed several methods to derive asymptotic or bootstrap confidence intervals and has compared their width and coverage level on the basis of simulations. Some of the possible approaches are presented and illustrated below on UK mercury data. The proposed methods to derive confidence intervals may be classified into three categories.

The first approach uses the central limit theorem and allows to derive asymptotic confidence intervals at level $1-\alpha$ for any parameter of interest θ estimated by a consistent estimator $\hat{\theta}$. It is defined by :

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{\mathrm{var}}(\hat{\theta})}.$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ percentile of the standard normal distribution and $\hat{\mathrm{var}}(\hat{\theta})$ is an asymptotic estimator of the variance of $\hat{\theta}$.

For example, a confidence interval for the empirical regional mean $\tilde{m}_R$ (see Section 4.3) can be derived using as variance :

$$\hat{\mathrm{var}}(\tilde{m}_R) = \sum_{i=1}^{r} w_i^2 p_i^2 \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

where censored x_{ij} have been replaced by an imputed value as a random uniform number between 0 and δ_{ij} (note that this is the only imputation method that makes sense when calculating confidence intervals). Note that, this formula like the following ones takes the hypothesis of independence between locations and observations within each location, which is certainly hazardous but is a first possible approximation when time and position information are not available.

For the parametric regional mean $\hat{m}_R$ (see Section 6.2) based on complete or summary data, the variance will be derived as :

$$\hat{\mathrm{var}}(\hat{m}_R) = \sum_{i=1}^{r} w_i^2 \hat{\mathrm{var}}(\hat{m}_i).$$

Since $\hat{m}_i$ is a function of the maximum likelihood estimator of the parameters of the underlying local distribution (e.g. for the lognormal, $\hat{m}_i = \exp(\hat{\mu}_i + \hat{\sigma}_i^2/2)$), $\mathrm{var}(\hat{m}_i)$ might be estimated using the delta method. In the presence of censored data, the derivation of the variance of local parameter estimators needed in the delta method is not straightforward. Asymptotic formulae exist in the lognormal complete case (Hald (1949)) but Lecoutre suggests to estimate small sample variances by Monte Carlo simulations, a method which may be adapted to summaries and to any class of local distribution.

The central limit theorem may also be used to derive confidence intervals for the empirical quantiles $\tilde{x}_p^R$ as follows :

$$\tilde{x}_p^R \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n\tilde{f}_R^2(\tilde{x}_p^R)}}.$$