UCL

Institut de statistique

Consulting report N° 02-01

Report on:

Assessment of the concentration level of chemical substances in river networks

Part IV

Calculation of confidence intervals for regional parameters of interest

Client: Eurochlor - Brussels
Authors: Bernadette Govaerts, Eric Lecoutre, Philippe Vanden Eeckaut,
Status : Public

Institut de statistique, Université catholique de Louvain – Voie du Roman Pays, 20
B1348 Louvain-la-Neuve – Tel : (32)010-47.43.14 – Fax : (32)010-47.30.32 – http://www.stat.ucl.ac.be

Assessment of the concentration level of chemical substances in river networks

Part IV

Calculation of confidence intervals for regional parameters of interest

Eric Lecoutre
Bernadette Govaerts
Philippe Vanden Eeckaut
Institut de Statistique
Université Catholique de Louvain

Summary

The Institute of Statistics of UCL has developed during the year 99, in the framework of a contract with Eurochlor, a methodology to summarize the concentration level of chemical substances in the rivers of a region on the basis of monitoring data (Govaerts *et al* [2001]). The main idea of the approach is to build a statistical distribution to summarize the concentrations observed in the given region and to derive then from it any statistics of interest like a regional mean, percentiles... This methodology is essentially descriptive in the sense that no incertitude measure is attached to the estimators of the regional parameters.

This paper, which summarizes the related master thesis of E. Lecoutre [2001], proposes some solutions to derive confidence intervals for regional parameters of interest. Three general methods are investigated : confidence intervals based on the central limit theorem, bootstrap confidence intervals and non-parametric confidence intervals based on ranks. The three methods have been adapted to be able to derive confidence intervals for the region mean and percentiles when they are estimated through one of the possible estimation method : empirical method, and parametric method for complete and for summarized data.

The relative qualities of the proposed intervals are compared through a large simulation study and the limitations of the methodology are finally discussed.

1. Introduction

The Institute of Statistics of UCL has developed during the year 99, in the framework of a contract with Eurochlor, a methodology to summarize the concentration level of chemical substances in the rivers of a region on the basis of monitoring data. The main idea of the approach is to build a statistical distribution to summarize the concentrations observed in the given region and to derive then from it any statistics of interest like a regional mean, percentiles... The method has already been applied to several chlorine substances of the COMMPS data base (see COMMPS (1998) and Govaerts *et al* (1999)).

Three different approaches have been proposed to estimate such a statistical distribution : 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. The empirical method is illustrated in Figure 1.

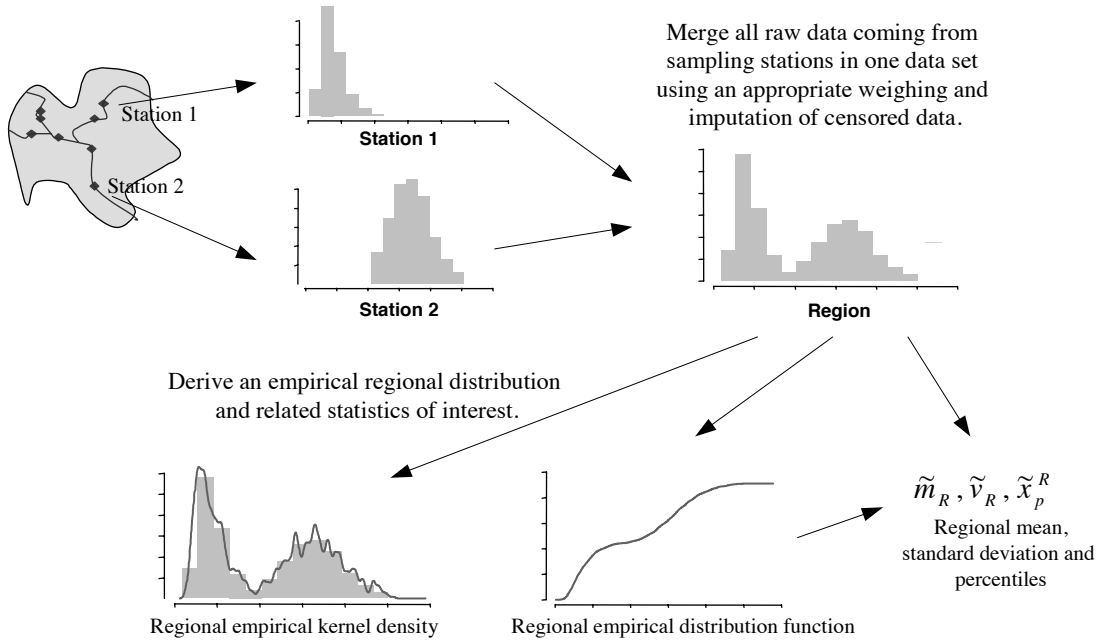

Figure 1 : Empirical method to derive a regional distribution

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. 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 2.

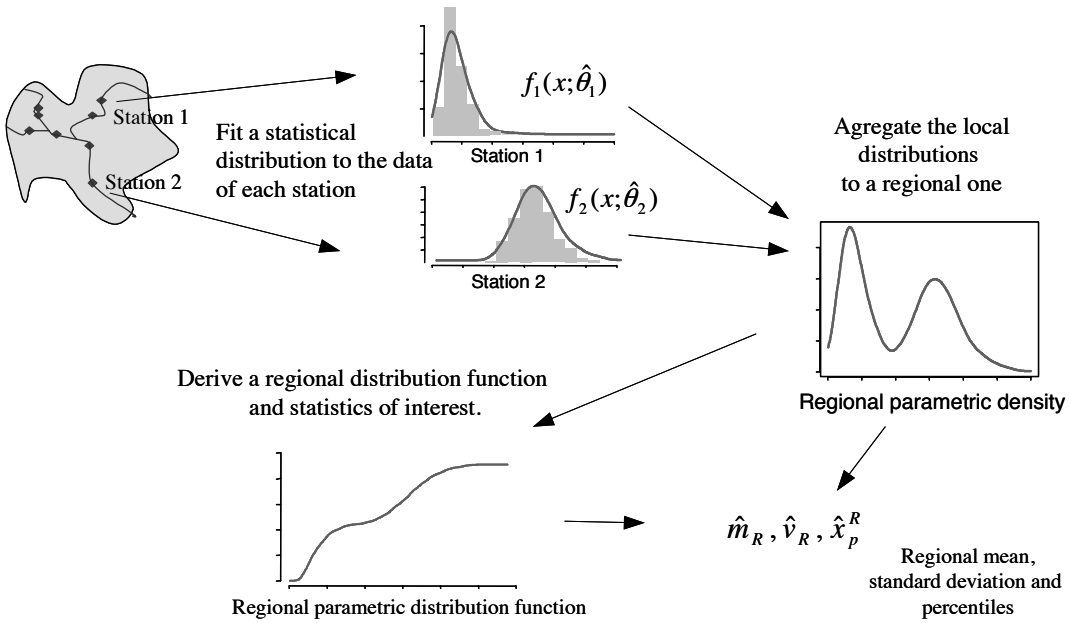

Figure 2 : Parametric method to derive a regional distribution

A detailed presentation of the developed methodology may be found in Govaerts *et al* [2001].

These different approaches proposed to summarize local data at a regional level are purely descriptive in the sense that, in the estimation of a regional mean of percentiles, nothing is said about the precision of the estimation. This is the purpose of inferential statistic which provides a theoretical framework to derive confidence intervals for the parameters of interest on the basis of estimated ones.

The goal of this paper is to propose some possible solutions to derive confidence intervals for the regional mean or percentiles (p0.5, p0.9...) from the estimated mean and percentiles obtained through one of the three possible approaches : empirical, parametric based on all data and parametric based on summary data.

This work has been realized by E. Lecoutre (2001) in the framework of his master thesis in statistics at UCL and the reader should refer to this text for the details not provided here.

This paper is organised as follows :

Section 2 discusses the inferential framework linked to the problem and the possible sources of imprecision of the estimated regional parameters. Section 3 presents first a brief review of necessary formulae and notations and discusses then possible ways to obtain confidence intervals for the regional parameters. Section 4 analyses, through simulations, the properties of possible intervals and compares the relative qualities of the different approaches regarding the estimation of each parameter of interest. Finally, section 5 gives some words of conclusion and discusses the applicability of the methods provided.

2. Inferential framework, sources of uncertainty

2.1 Inferential framework of the project

The classical inferential statistical methods typically suppose that a population of interest is studied through a set of data observed on a sample extracted from the population. In the context of the project, the population of interest may be seen as the total quantity of water present in a river network over a certain period of time. A good image is to see this water as a big lake from which small water samples are extracted. Each possible sample of water in this lake should have the same chance to be drawn in the sample. The samples are in our case not drawn randomly from the population but taken in imposed sampling station at chosen times along the period of interest.

A typical goal in inferential statistics is then to try to estimate the value of an (unknowm) parameter of interest (say θ) of the population from the data sample. In our context, a parameter of interest may be the mean concentration of a substance of interest. Such theoretical mean concentration should be seen as the arithmetic mean one would obtained from an infinite number of measurements made without error at random places and at random time over the river network.

In practice, the parameter of interest θ will be estimated using an "estimator" defined as a mathematical function of the observations of the data sample. For a given parameter of interest, several estimators may usually be derived as in our context where three methods are proposed to estimate the regional mean (empirical, parametric on all or summarized data). Moreover, for a given estimator, two data samples will of course give two different numerical values for the estimation.

The inferential problem is then "how close to the true parameter of interest is an estimated value ?". The classical approach to answer such a question consists of providing a "confidence interval" around the estimated value which has a big probability to contain the unknown "true" parameter. This confidence interval should ideally take in account all the possible sources of uncertainties linked to the parameter estimation.

From this small discussion, it appears clearly that the construction of such intervals necessitates to define first precisely what is the population from which the sample is drawn and second to list as completely as possible the sources of uncertainty of the estimation.

In his work, E. Lecoutre gives two possible definition of the population of interest for a given river network. The first supposes that, even if the measurements are only made in a given set of sampling stations, they are representative of the whole river network over the period of interest. The interpretation will be referred below as "random design" and is the more intuitive one. A more restricted approach defines the population as the water passing over the period of time through the sampling stations only. This second possible interpretation of the population will be referred later as "fixed design".

The methods proposed to build confidence intervals will be different if one wants to estimate parameters of interest for the first or second (more restricted) population.

2.2 Sources of uncertainty

Three categories of sources of uncertainties in the estimation of the regional parameters may clearly be identified in our context : statistical methods used, sampling, and measurement uncertainty.

2.2.1 Statistical methods used

The quality of a parameter estimator will first depend on the techniques used to derive the estimator formula and the statistical hypothesis taken on the data. More precisely in our context the estimator's quality relies on the following factors :

- The general method chosen to derive the regional distributions : empirical and parametric (on all or summarized data). Many other techniques could have been envisaged.
- In the parametric method, the statistical distribution chosen to fit local data (lognormal, gamma...) and the method used to estimate the local distributions parameters (maximum likelihood, log normal probability regression...).
- The method used to deal with censored data (e.g. imputation in empirical distributions and adaptation of ML in the parametric approach).
- The independence hypothesis taken on the data. The proposed methodology supposes that the data are independent from one station to another and from time to time in a same station. This is certainly not true but leads to significant simplifications in the development of estimation formulae. The intuition is that the independence hypothesis does not have a big impact on the quality of the estimators but may affect greatly the quality of the confidence intervals.

Typically, in classical statistical inference methods, one supposes that the statistical hypothesis taken to model the data are correct and it is the purpose of robustness analysis to verify if the results remain acceptable even if some hypothesis are not perfectly verified. The work of E. Lecoutre follows this usual approach. All the confidence intervals have been developed supposing the independence of the data between and within stations and that the local data follow a given distribution in the parametric approach.

2.2.2 Sampling and representativeness of the sample

Inferential statistics supposes in theory that data used in a parameter estimation are coming from a random sample extracted from the population or at least from a "representative" sample from the population. In our context, this is certainly never verified due to practical organizational problems and one should at least expect :

- that stations are "well spread" over the river network,
- that water samples are taken regularly over the sampling period and
- that the method used to take water samples from the river gives a representative view of the real situation at that point on the river (deepness...).

When the first hypothesis is far to be verified, the methodology developed in Govaerts *et al* (2001) suggests to attach a weight to each station depending of the part of the river network that the station is supposed to represent (see Govaerts *et al* (2002)).

Note that, even in such ideal case, the sample will remain a subset of the population and parameter estimations will always be uncertain. This is the classical purpose of statistical inference. One supposes that the sample is representative, that the statistical methods chosen are correct, that the data are observed without error and one tries to get in a such ideal situation, an idea of what is the value of a theoretical parameter.

2.2.3 Measurement uncertainty

Substance concentrations measured in water samples are obtained through a chain of manipulations which are at each stage subject to uncertainties. In the same station, with the same instrument, method and operator, two measurements of the same water sample never give identical measured values. Such variation is commonly called repeatability. In a same station and a fortiori between stations, measurement devices, techniques and methods may vary and interlaboratory tests often show systematic bias between stations and very different measurement fidelity. An additional difficulty comes from the fact that many measured concentrations lie below the determination limit of the measurement device and are then not available for the statistical analysis (data are said to be censored).

All these factors of uncertainty may significantly affect the quality of the data in our context because, for many products, the concentrations are very low (and often below the determination limit) and observed concentrations have very low relative precision.

The work of E. Lecoutre does not at all attack this aspect of the problem. This means in practice that the confidence intervals may only be applied if the measurement uncertainties may be considered as low compared to the variation of the concentration over the region of interest.

3. Possible methods to build confidence intervals

3.1 Necessary notations and formulae

Formal notation is necessary to introduce possible confidence intervals. All details on these notations and formulae are developed and commented in Govaerts *et al* [2001].

General notations :

- R : the region of interest.
- r : the number of sampling stations in the region R.
- S_i : the i^{th} station in region R.
- m_R and x_p^R : the theoretical regional mean and percentile of order p for which confidence intervals have to be developed.
- $f_R(x)$ and $F_R(x)$: the theoretical regional density and distribution function of the concentration.
- n_i : number of measurements made in station S_i over the sampling period.

x_{ij} : the j^{th} measured concentration in station S_i. If the data is censored, δ_{ij} is the corresponding detection limit.

N : the total number of measurements in the region over the sampling period.

w_i : the weight allocated to station S_i in the regional aggregation. Details on a methodology to choice such weights may be found in Govaerts *et al* [2002]. An easy default weighting is $w_i = 1/r$.

Notations related to the empirical approach to derive regional statistics.

$\tilde{m}_R$ is the empirical regional mean calculated as : $\tilde{m}_R = \sum_{i=1}^{r} w_i \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

$\tilde{x}_p^R$ is the empirical regional percentile of order p calculated as

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

where $\tilde{F}_R = \sum_{i=1}^{r} w_i \frac{1}{n_i} \sum_{j=1}^{n_i} I(x_{ij} \leq x)$ with $I(x_{ij} \leq x) = 1$ if $x_{ij} \leq x$ and 0 otherwise.

Notations related to the parametric approach

$\hat{m}_R = \sum_{i=1}^{r} w_i \hat{m}_i$ is the estimated parametric regional mean where $\hat{m}_i$ is the mean in station S_i estimated from a local parametric density (e.g. lognormal).

$\hat{x}_P^R = \min\{x \in R^+ \mid \hat{F}_R(x) \geq p\}$ is the estimated parametric regional percentile of order p calculated on the basis of the regional distribution function defined as the weighted sum of local distribution functions.

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

Different methods are now proposed to derive confidence intervals for m_R and x_p^R from the estimated ones : $\tilde{m}_R$, $\tilde{x}_p^R$, $\hat{m}_R$ or $\hat{x}_P^R$

3.2 Confidence intervals based on the central limit theorem

The first approach proposed uses the central limit theorem to derive asymptotic confidence intervals at level 1-α for any parameters of interest θ estimated by a consistent estimator $\hat{\theta}$. Such interval is defined by :

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

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ percentile of the standard normal distribution and $\hat{\text{var}}(\hat{\theta})$ is an asymptotic estimator of the variance of $\hat{\theta}$. Such interval is only valid under certain regularity conditions and when the sample size is large. Small sample properties may be investigated through simulations. In our context the important practical difficulty is to calculate $\hat{\text{var}}(\hat{\theta})$.

Lecoutre [2001] has applied the central limit theorem to derive confidence intervals for the empirical and parametric regional means and empirical regional percentiles. These three

intervals are only valid in the fixed design framework where one supposes that the inference is limited to the sampling stations and not to the whole river network.

For the empirical mean, the variance of $\tilde{m}_R$ may be estimated using the following formula :

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

In this formula, the observations below the detection limits are replaced by an imputed value. It is recommended to use an uniform imputation approach on the interval [0,DL] because the imputation with DL or DL/2 will have for effect to reduce artificially the estimated variance and, in consequence, the length of the confidence intervals.

For the empirical percentiles, the following formula due to Kendall may be used to estimate asymptotically the variance of $\tilde{x}_p^R$:

$$\hat{var}(\tilde{x}_p^R) = \frac{p(1-p)}{N \tilde{f}_R(\tilde{x}_p^R)}$$

where $\tilde{f}_R(\tilde{x}_p^R)$ is the empirical density calculated in $\tilde{x}_p^R$.

For the parametric mean (based on all or summarized data). The variance will be derived as :

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

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, $\hat{m}_i = \exp(\hat{\mu}_i + \hat{\sigma}_i^2/2)$), $\hat{var}(\hat{m}_R)$ may be estimated using the delta method. Unfortunately, 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 is only valid for constant detection limits. 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.

3.3 Bootstrap confidence intervals

A second alternative is the bootstrap approach which consists in redrawing with replacement several (*B* say) samples from the original data set, re-estimate the parameter of interest θ from each bootstrap sample $\hat{\theta}_i$ and derive the confidence intervals from the observed quantiles of the *B* bootstrap estimators (Davison and Hinkley (1997)):

$$\left[q_{\alpha/2}(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B), q_{1-\alpha/2}(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B) \right]$$

Bootstrap may be easily applied to derive confidence intervals for moments and percentiles in the empirical or parametric case when data are not summarized. Lecoutre presents three possible approaches to derive bootstrap confidence intervals :

- In the first approach, called *simple boostrap*, the bootstrap samples are drawn directly from the whole data set using a weight w_i / n_i for each original observation to take into account that each station does not have the same importance in the region and that the number of observations is not constant from station to station.

- The second approach, called *two level boostrap,* takes into account the fact that the data are organised in two levels : r stations in the region and n_i observations within each stations. The samples are drawn in two stages, first, r stations are drawn (with replacement) in the r possible stations and then n_i observations in each chosen station. This framework respects the random design scheme were stations are supposed to be only a sample of possible stations in the region. The station sampling must be done while taking into account the respective weights w_i attached to each station. In the fixed design framework, an alternative to this method consists of keeping the r original stations in the bootstrap sample and draw only within the stations.

  Note that in this resampling, censored data are resampled as other data and treated adequately afterwards by the chosen estimation method (empirical or parametric).

- When data are summarized, resampling within stations becomes impossible since the data are not available and Lecoutre proposes to replace it by monte carlo simulations of local parameters in the neighbourhood of estimated ones assuming an asymptotic normal distribution for the local estimators. This third approach, called *semi-parametric bootstrap*, can be applied both to summarized and complete data and has the advantage to be less computer intensive than former bootstrap methods.

3.4 Non parametric confidence intervals for percentiles

A last approach may be used to derive confidence intervals for empirical percentiles on the basis of ranks of observed data (Helsel and Hirsh (1992)). If $\tilde{x}_p^R$ is the percentile of interest, the limits of the confidence interval are defined by the two observations of the whole ordered sample of data available for the region with ranks :

$$Np \pm z_{1-\alpha/2}\sqrt{Np(1-p)}$$

These ranks must be adapted to take into account that each observation in the region does not have the same weight in the definition of the regional statistics. This approach is only valid in the fixed design framework.

3.5 Summary of available methods

The previous section shows possible ways to derive confidence intervals for regional mean and percentiles estimated through an empirical or parametric approach. All methods can not be applied to all parameters of interest and even for some estimators no possible confidence interval is provided. Moreover, the method to derive a confidence interval depends on the target population denoted as fixed and random design in section 2.

Table 1 gives a summary of the list of possible confidence intervals proposed for the regional parameter we are interested in : the mean and percentiles. The possible intervals will depend on the way the parameters are estimated (empirical, parametric on all data and

parametric on summarized data). Moreover, each type of confidence interval will be valid either in the fixed or/and in the random design framework.

Class of confidence interval	Regional Mean						Regional Percentile					
	Empirical		Param All		Param Sum		Empirical		Param All		Param Sum	
	F	R	F	R	F	R	F	R	F	R	F	R
Central limit theorem	X		X		X		X					
Simple bootstrap	X	X	X				X	X				
Two level bootstrap	X	X	X*	X*			X	X	X*	X*		
Semiparametric bootstrap			X	X	X	X			X	X	X	X
Non parametric							X					

Table 1: List of confidence intervals proposed in Lecoutre [1991]. X* have not been tested by simulations.

4. Confidence intervals validation and comparison simulations

4.1 Simulations goals

The validity of the confidence intervals derived in section 3 for regional parameter of interest has been tested by Lecoutre [2001] through a huge set of simulations. These simulations were aimed to verifying whether the confidence intervals based mostly on asymptotic results are still valid when the sample size is limited and whether the very general theoretical results used in their construction are applicable in the presence of censored data and within a non common set of data structured in two levels : stations and data within stations.

These simulations will allow to verify if two crucial properties are close to be verified by the proposed confidence intervals. First, a confidence interval at a level of confidence (1-α) is supposed to contain, on average, the true value of the target parameter 100(1-α)% of the cases. Moreover, the confidence interval should be as short as possible around the target value. The "coverage level" and interval length have been measured for the different possible intervals in a wide variety of situations.

4.2 Simulations principles

The general principle of a simulation study in such context is to simulate data for which one knows what are the (usually unknown) values of the parameters to estimate. This is the only way to verify at the end if the confidence intervals cover the target values. The parameters of interest tested by Lecoutre [1991] are : the regional mean m_R and the 50th, 90th, 95th and 99th regional percentiles : $x_{0.5}^R, x_{0.9}^R, x_{0.95}^R$ and $x_{0.99}^R$.

A second principle consists of simulating the data with the same statistical hypothesis taken in the construction of the confidence intervals. In his simulations Lecoutre [2001] used two main properties : firstly the independence of the data between and within sampling stations and secondly, the lognormality of the data within each sampling station. Nothing has been done on the robustness of the methods if these hypothesis are not verified (which is clearly the case for the independence).

The basis of the simulation is the region of interest. Lecoutre built two possible regions of interest : one with a wide variety of sampling stations (that is a large possible set of values for the parameters μ and σ of the corresponding lognormal distributions) and one region with quite homogeneous sampling stations. The possible values for the sampling station parameters were chosen on the basis of real data sets of the COMMPS data base.

In these two regions, the "true" parameters of interest, $m_R, x_{0.5}^R, x_{0.9}^R, x_{0.95}^R, x_{0.99}^R$ were calculated. The simulations consisted then in drawing randomly data in the region, estimate the parameter of interest and related confidence intervals and quantify their quality with respect to the expected parameters through their length and coverage level.

More precisely, in the random design framework, a single simulation works as follows :

1. Draw a set of r stations in the region (that is a set of pairs (μ_i, σ_i)).
2. For each sampling station S_i, draw n_i observations distributed independently as a lognormal(μ_i, σ_i) distribution.
3. Censor all data which are below a given detection limit DL (this means that in the region we suppose that all the instruments have the same detection limit)
4. Calculate summarized data within each station.
5. Estimate the parameters of interest $m_R, x_{0.5}^R, x_{0.9}^R, x_{0.95}^R, x_{0.99}^R$ with the three possible approaches (empirical (with uniform imputation), parametric using all the data, parametric with summary data).
6. Calculate for each of the (15=5x3) estimated parameters the possible 95% confidence intervals (see table 1).
7. Verify for each confidence interval if the true parameter belongs to it and calculate the length of the interval.

For each given set of values (r, n and DL) Lecoutre repeated this single simulation 500 times to get a good idea of the performance of each intervals in a given situation. Several combinations of r, n and DL were tested : r = 10 to 50, n = 10 to 50 and DL = 0.05 and 0.1 which gave in practice from 5 to 40% of censored data.

In the fixed design framework, the same structure of simulation was used but the reference parameters $m_R, x_{0.5}^R, x_{0.9}^R, x_{0.95}^R, x_{0.99}^R$ were no longer calculated at a regional level but on each given set of r stations drawn in the region.

4.3 Simulation results structure

The following sections discuss the performances of the confidence intervals for three regional parameters : the regional mean (in section 4.4), the regional median (in section 4.5) and the regional 95th percentile (in section 4.6). In each section, a first paragraph summarizes the general conclusions for the parameter of interest and results are then detailed for the three different estimation approaches : empirical for raw data, parametric for all data and parametric for summary data.

For those detailed results, the simulation results for r=25 and n=25 are presented in a summary table. For each possible estimation method, the performances of the possible confidence intervals are given both for the fixed (F) and random (R) design frameworks.

Two measures are proposed to evaluate the performance : the coverage rate (percentage of intervals which contain the true parameter value) and the average length of the intervals.

Some additional comments give an interpretation of the results and highlight the most desirable intervals. When useful, the discussion is extended to simulation results for other r and n (full numerical results can be found in Lecoutre [2001]).

4.4 Comparison of confidence intervals for the regional mean

In the fixed design framework, the intervals based on the central limit theorem (CLT) should be recommended. They ensure high cover rates but have quite high lengths. This length can be explained by the fact that CLT provides symmetrical intervals, where as the regional mean estimators have asymmetric distributions. Consequently, to cover the real interval, a symmetric interval has to be longer.

As the random design is highly preferable, in terms of interpretability and realism, we encourage to use methods adapted to this design. In this framework, the bootstrap is the only solution and is computational intensive.

We particularly retain the semi-parametric bootstrap, which doesn't work for the fixed design, but ensures correct cover rates for the random design. Even if this good cover rate is due to excessive lengths, a correct cover rate is so desirable that we will tend to advocate this security approach (adopt those worth-case bounds).

As a general result, the simulations show that increasing r, the number of stations, is always preferable that increasing n, the number of observations per station.

4.4.1 Empirical estimation of the regional mean

	$r=25$, $n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	CLT	90	0.183	89	0.190	93	0.039	93	0.040
	CLT agg.[1]	**93**	**0.197**	**92**	**0.204**	**93**	**0.040**	**93**	**0.040**
	Simple bootstrap	92	0.202	92	0.209	93	0.039	92	0.040
	Two-level bootstrap	88	0.183	90	0.188	89	0.038	*18*	0.037
R	**Simple bootstrap**	**89**	**0.416**	**91**	**0.437**	**97**	**0.060**	**97**	**0.059**
	Two-level bootstrap	90	0.411	90	0.430	97	0.059	97	0.059

In the fixed design framework, the confidence intervals based on the CLT should be preferred since they give acceptable properties and can be derived quickly. Nevertheless, their default is to provide symmetric intervals when the distribution of the empirical mean appears to be asymmetrical in the simulations. Simple bootstrap gives also good intervals but is much more computer intensive than the CLT approach.

[1] This CTL interval is based on summarized (or aggregated) local data.

For the random design the two bootstrap approaches seem to be quite equivalent.

The difference of interval lengths for the two sampling designs, even for the homogeneous region 2, lead us to recommend a careful use of the fixed design intervals : the random design is closer to reality, taking into account that the stations are not necessary representative of the region.

4.4.2 Parametric estimation of the regional mean on complete data

	$r=25, n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**CLT**	**100**	**0.593**	**100**	**0.552**	**100**	**0.142**	**100**	**0.079**
	s.-param. bootstrap	91	0.310	91	0.323	*73*	0.068	83	0.063
R	**s.-param. bootstrap**	**96**	**0.609**	**97**	**0.630**	**97**	**0.109**	**99**	**0.102**

For the fixed design, we highly recommend CLT intervals, even if their high coverage rates have to be counterbalanced by a high interval lengths. Nevertheless, this is a necessary cost, as the semi-parametric bootstrap shows low performances.

In the random design framework, the semi-parametric bootstrap is the only candidate, but reveals itself a good candidate. Theoretical cover rates are reached, with acceptable lengths compared with others methods.

4.4.3 Parametric estimation of the regional mean on summarized data

	$r=25, n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**CLT**	**100**	**0.616**	**100**	**0.581**	**100**	**0.148**	**100**	**0.086**
	s.-param. bootstrap	*67*	0.296	*73*	0.344	68	0.073	69	0.074
R	**s.-param. bootstrap**	**90**	**0.576**	**96**	**0.679**	**97**	**0.119**	**99**	**0.119**

For the summary data approach, results are quite similar to the previous approach, yielding the same practical recommendations. Note that the interval lengths obtained for complete and summarized data are very similar which is a good point in favour of the use of summarized data.

4.5 Comparison of confidence intervals for the regional median

For the regional percentiles, fewer methods are available than for the regional mean, especially for the parametric approach but one can still exhibit methods with a high performance. For the parametric approaches, the semi-parametric bootstrap is the only method available. It gives good performances but is computer intensive.

Thus, for the median, we really encourage to estimate the regional median on the basis of raw data (empirical approach) and use non parametric confidence intervals which are very easy to derive.

Note that for the parametric approaches, increasing n alone allows to decrease the length of the interval, where as increasing r tends to reduce the observed bias (this for summary data).

4.5.1 Empirical estimation of the regional median

	$r=25, n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**Non parametric**	**99**	**0.068**	**99**	**0.063**	**95**	**0.021**	**96**	**0.020**
F	CLT	100	0.099	100	0.102	98	0.025	98	0.025
F	Simple bootstrap	98	0.066	98	0.062	94	0.020	93	0.019
F	Two-level bootstrap	94	0.052	93	0.047	93	0.019	93	0.019
R	**Simple bootstrap**	**98**	**0.031**	**97**	**0.059**	**98**	**0.031**	**98**	**0.029**
R	Two-level bootstrap	99	0.031	97	0.059	99	0.031	98	0.029

For the empirical estimation of the regional median, all the available methods seem to give acceptable results.

For the fixed design, we definitively prefer the non parametric approach, as this is an easier method to program, and also the fastest one (one just has to sort the data).

For the random design, both bootstrap methods seem to work. If one is interested to derive confidence intervals for the empirical mean and the median, the simple bootstrap should be preferred because the two-level bootstrap is not efficient for the mean. A same set of bootstrap simulations can then be used to derive simultaneously confidence intervals on both parameters.

4.5.2 Parametric estimation of the regional median on complete data

	$r=25, n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**s.-param. bootstrap**	**100**	**0.069**	**99**	**0.064**	**96**	**0.022**	**97**	**0.022**
R	**s.-param. bootstrap**	**100**	**0.038**	**99**	**0.103**	**100**	**0.038**	**100**	**0.036**

For the parametric approach on whole data, only the semi-parametric bootstrap is available and it seems to give good results. Theoretical cover rate are reached and the length are acceptable, compared to those obtained with empirical methods.

4.5.3 Parametric estimation of the regional median on summarized data

	$r=25, n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**s.-param. bootstrap**	**93**	**0.076**	**95**	**0.065**	**98**	**0.022**	**97**	**0.022**
R	**s.-param. bootstrap**	**100**	**0.038**	**99**	**0.119**	**100**	**0.038**	**100**	**0.039**

For the parametric approach on summary data, the results are more or less the same, that is acceptable ones for semi-parametric bootstrap. Nevertheless, it appears that using summary data may lead to some bias in the median estimation, particularly when r is low. Increasing

n allows to reduce the length of the interval, as expected, where as increasing r only allows to reduce the bias even if the total number of observations increases.

4.6 Comparison of confidence intervals for the regional 95th percentile

For the 95th regional percentile, a parameter of great interest, the same methods are available than for the median (which is the 50% percentile). It appears that parametric approaches imply a bias in the regional estimator when r is low, so that intervals are valid only when considering regions with a high number of stations (say $r > 100$).

As the statistic of interest has a high variance and is less stable, there is a higher sensitivity to the considered sampling design with the practical implication that the intervals will be really too short for the fixed design. Thus, one should consider random design for this percentile, which implies the use of bootstrap a computer-intensive method (specially as r should be high...).

4.6.1 Empirical estimation of the regional 95th percentile

	$r=25$, $n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**Non parametric**	**96**	**0.831**	**93**	**0.895**	**95**	**0.203**	**95**	**0.203**
F	CLT	*86*	0.635	86	0.682	91	0.179	90	0.179
F	Simple bootstrap	94	0.793	94	0.824	94	0.195	93	0.196
F	Two-level bootstrap	92	0.727	92	0.786	94	0.191	93	0.191
R	**Simple bootstrap**	**91**	**1.712**	**93**	**1.793**	**99**	**0.292**	**97**	**0.285**
R	Two-level bootstrap	90	1.681	92	1.741	98	0.289	96	0.281

As for the median, we bring up the non parametric interval for the fixed design and the simple bootstrap for the random design. For the less homogeneous region (region 1), the difference between fixed and random design appears to be important, so that we recommend to use the bootstrap in the random design framework.

4.6.2 Parametric estimation of the regional 95th percentile on complete data

	$r=25$, $n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**s.-param. bootstrap**	**96**	**0.924**	**95**	**0.961**	*71*	**0.241**	**85**	**0.228**
R	**s.-param. bootstrap**	**97**	**2.040**	**98**	**2.106**	**97**	**0.400**	**99**	**0.376**

As for the median, semi-parametric bootstrap gives acceptable results, with a problem of bias when r is low. Obtained lengths are more or less comparable to those obtained with an empirical approach.

4.6.3 Parametric estimation of the regional 95th percentile on summarized data

	$r=25$, $n=25$	Region 1				Region 2			
		DL=0.05		DL=0.10		DL=0.05		DL=0.10	
	Method	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**	**Cover.**	**Length**
F	**s.-param. bootstrap**	*55*	0.905	*68*	0.977	*77*	0.239	**83**	0.233
R	**s.-param. bootstrap**	**95**	1.757	**96**	2.003	**98**	0.418	**99**	0.402

Once again, semi-parametric methods give similar results for complete and summarized data. The bias is more present especially is the fixed design framework .

5. Conclusions and domain of application of the confidence intervals

This paper (and the related master thesis of E. Lecoutre [2001]) attempts to propose some solutions to derive confidence intervals for regional summaries of the concentration level of a chemical substance in a river network. The properties of the proposed methods have been studied through a simulation study which show that, in the framework of the simulations, the intervals qualities are very acceptable. For each parameter of interest at least one method gives good coverage rates and acceptable interval lengths.

These nice results are nevertheless not really applicable in practice for several reasons :

First, the simulations have been done under precise statistical hypothesis which are not necessarily valid in practice. They suppose, for example, that the data are independent from station to station and within each station. This hypothesis is certainly not true and one should test the impact of non independence on the interval properties. When high positive autocorrelation is present, proposed confidence intervals are probably too optimistic (too short).

Second, the intervals take only in account the statistical sampling incertitude due to the fact that only a part of the "population" (river network) is observed over the period of interest. They do not cover other sources of uncertainty (see section 2.2) and possible lack of quality of the data.

Third, the calculation of an appropriate confidence interval is not always easy. In fact, when considering the random design - and we think this is the one that makes sense in this context – the bootstrap is the only method available. This technique seems to give acceptable results but is too computer intensive to be used on a daily basis. Moreover, for parametric approaches, the reduction of a possible bias requires to have at our disposal a high number of gauging stations ($r>100$), which increases drastically the computational time.

Note finally that most of the proposed techniques are not acceptable anymore when considering high percentiles (such as 99%). Those percentiles are so unpredictable that their estimation relies on a specific statistical area (extreme value theory).

So, as a final conclusion, we can not recommend the use of such confidence intervals in official reports. For the reasons given above, their adequacy to the context must still be investigated more deeply. We can nevertheless certainly encourage their use as an informative tool giving an idea of the minimum incertitude rate on the estimated parameters. When the purpose is risk assessment, we recommend to apply several of the available methods and to consider at a final stage the union of the different intervals (see Govaerts *et al* [2001] for an illustration on mercury data).

Bibliography

COMMPS (1998), *Proposal for a list of priority substances in the context of the draft water framework directive COM(97)49 FIN; Results of the COMMPS procedure*, Fraunhofer Institute for Environmental Chemistry and Ecotoxicology, Schmallenberg.

Davison, A.C. and Hinkley D.V. (1997), *Bootstrap methods and their application*, Cambridge university press.

Govaerts B. and P. Vanden Eeckaut (1999), *Aquatic toxicology and risk to the environment and/or the human health : spatial and trends analysis for monitoring data in European regions, a primer statistical and exploratory data analysis*, Consulting report, Institut de Statistique, UCL.

Govaerts B., B. Beck, E. Lecoutre, C. Lebailly and P. Vanden Eeckaut (2001), *From monitoring data to regional distributions: a practical methodology applied to water risk assessment*. Discussion paper 0127 of the Institute of Statistics UCL.

Govaerts B., E. Lecoutre and P. Vanden Eeckaut (2002), *Assessment of the concentration level of chemical substances in river networks, Part III, Choice of a weighing method in the aggregation of local monitoring data to a regional distribution*. Consulting report, Institut de Statistique, UCL.

Hard A. (1949), *Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point*. Skandinaviavisk Aktuarietidskrift.

Helsel, D.R. and Hirsh, R.M. (1992), *Statistical methods in water resources*, chapter 3, Elsevier.

Lecoutre E. (2001), *Concentration de produits chimiques dans les rivières d'une région à partir de données de monitoring. Intervalles de confiance autour de statistiques régionales*, Mémoire de DEA en statistique, UCL, Louvain-la-Neuve.

Stuart A. and Ord J.K. (1987), *Kendall's advanced theory of statistics*, volume 1. Charles Griffin and Compagny, 5th edition.