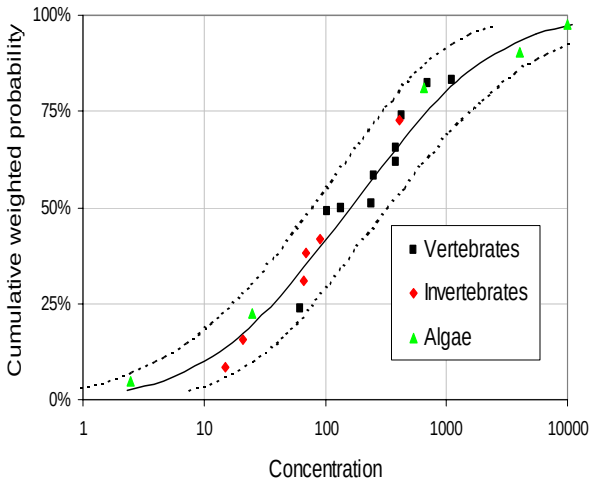
SPECIES SENSITIVITY WEIGHTED DISTRIBUTION SOFTWARE USER'S GUIDE

Version 1.0 – December 2003



Developed by Electricité de France (EDF)
With Institut de l'Environnement Industriel et des Risques (INERIS)
Methodology and design: C.Duboudin - EDF
Code development: R.Aletti - Simulog and C.Duboudin
Contacts: Ph.Ciffroy - EDF: philippe.ciffroy@edf.fr,
H.Magaud - INERIS: helene.magaud@ineris.fr

CONTENTS

INSTALLATION OF THE SSWD TOOL

What is the SSWD tool?

The SSWD procedure is a Macro ".xla" of Excel. It enables to build species sensitivity weighted distributions (SSWD) [11,16] and to calculate hazardous concentration (with confidence limits associated) for different thresholds of protection ($HC_{th\%}$ with different values of *th*). This SSWD tool is based on a paper published in the Environmental Toxicology and Chemistry Journal [17].

In this study, various methods of constructing a species sensitivity distribution for a same substance and a same sample of ecotoxicity data are compared. Each method is characterized by a different way of taking into account intra-species variation and weight of each trophic level (proportion of data in the vertebrates, invertebrates and algae), as well as by the statistical method of calculation of the HC_5. Those methods are tested on 15 substances using chronic NOEC data available in the literature and produce results which may be separated by one order of magnitude for the same substance. It turns out that weight of each taxonomic group or trophic level and intra-species variation have so much even more effect on the value of the HC_5 than the statistical method used to construct the distribution. So a procedure in which ecotoxicity concentration data can be weighted to account for redundant data for each species or genus and for the disproportion in the data number between the taxonomic groups or trophic levels was developed.

System requirements

Windows 95 or more.
Microsoft Excel 97 or more.

Warning: the **Complementary Macros** of Excel SOLVER.xla, ATPVBAEN.xla and FUNCRES.xla must be installed.

The decimal separator must be: "," or ".". No thousand's separator is accepted.

Procedure to install SSWD.xla

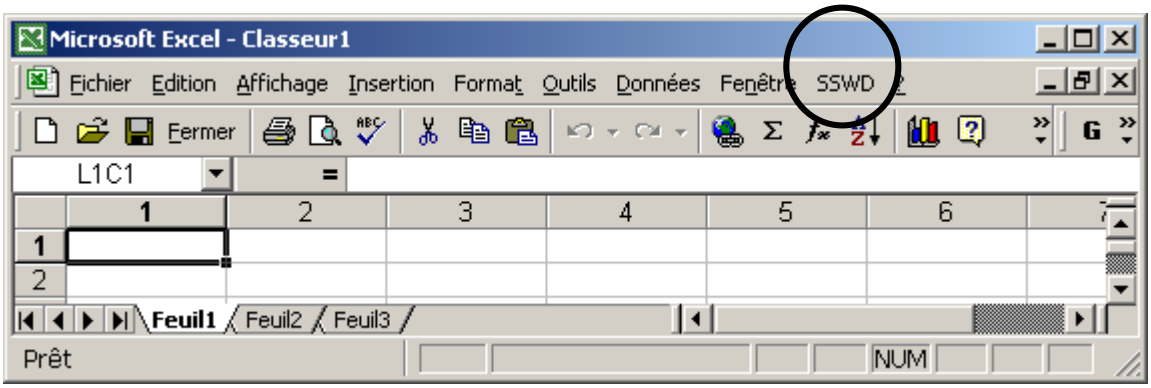
To install the **SSXD.xla** Macro in your system:

1. Put the **SSWD.xla** file in the chosen directory.

2. Open Excel.

3. From the main menu of Excel, choose **Tools ▶ Complementary Macros** to open the **Complementary Macros** dialog box. In this dialog box, select the options **Solver**, **Analysis Tools** and **Analysis ToolsVBA** if they are not yet selected. Click on the **OK** button to finish this operation. If these options are not proposed in the list of complementary macros, you must reinstall Excel with the specific option **Complementary Macros**. See the help of Excel for this.

4. Reopen the **Complementary Macros** dialog box and find with the **Browser** button the file **SSWD.xla** in your directory. The option **SSWD** appears now in the list of the options like **Solver**, **Analysis Tools** and **Analysis ToolsVBA**. Click on the **OK** button to finish the operation. A new menu **SSWD** appears now in the main menu of Excel. See below.



For uninstall the **SSWD.xla** Macro, reopen the **Complementary Macros** dialog box and deselect the option **SSWD** in the list of complementary macros. The **SSWD** option must disappear in the main menu of Excel. Before installing a new version of **SSWD.xla**, you must uninstall the old one.

USING SSWD PROCEDURE

Introduction

To use the SSWD procedure:

1. Open Excel.

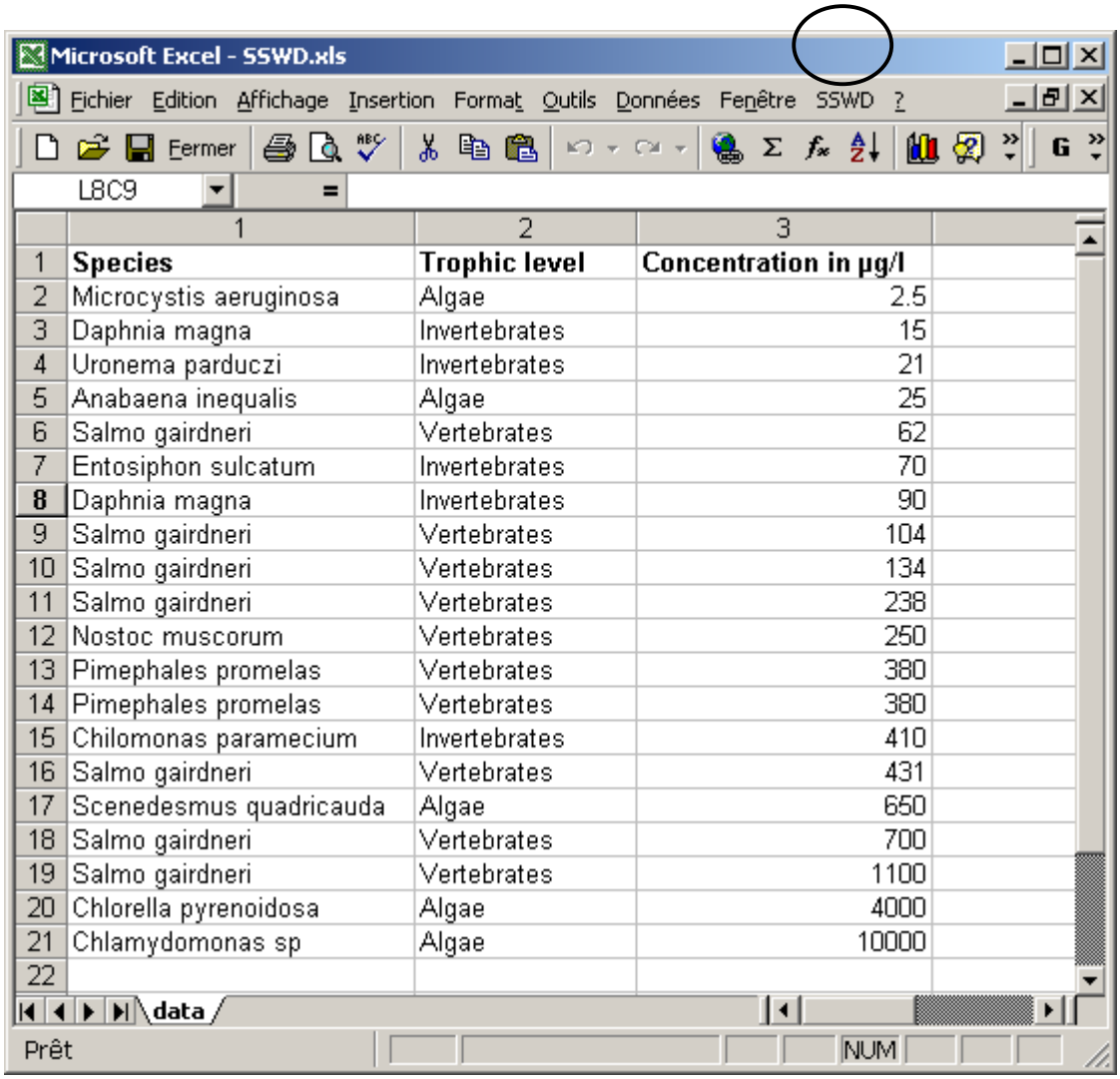
2. Put the database to be used in an Excel Worksheet.

   The data must be organized in columns with headings. A minimum of three columns is needed: (1) the column of the name of the species (or the genus) used in the ecotoxicology test; (2) the column of the trophic level (or the taxonomic group) of the species; (3) the column of the toxicological test results (concentration data). All the data can belong to the same trophic level (or taxonomic group), even to the same species. The number of taxonomic groups is not limited. Worksheet names do not contain any space. See example below. Be careful, the name of the Excel data worksheet must not contain the characters:

   ( ) ; , ! - # ‘ “ etc.

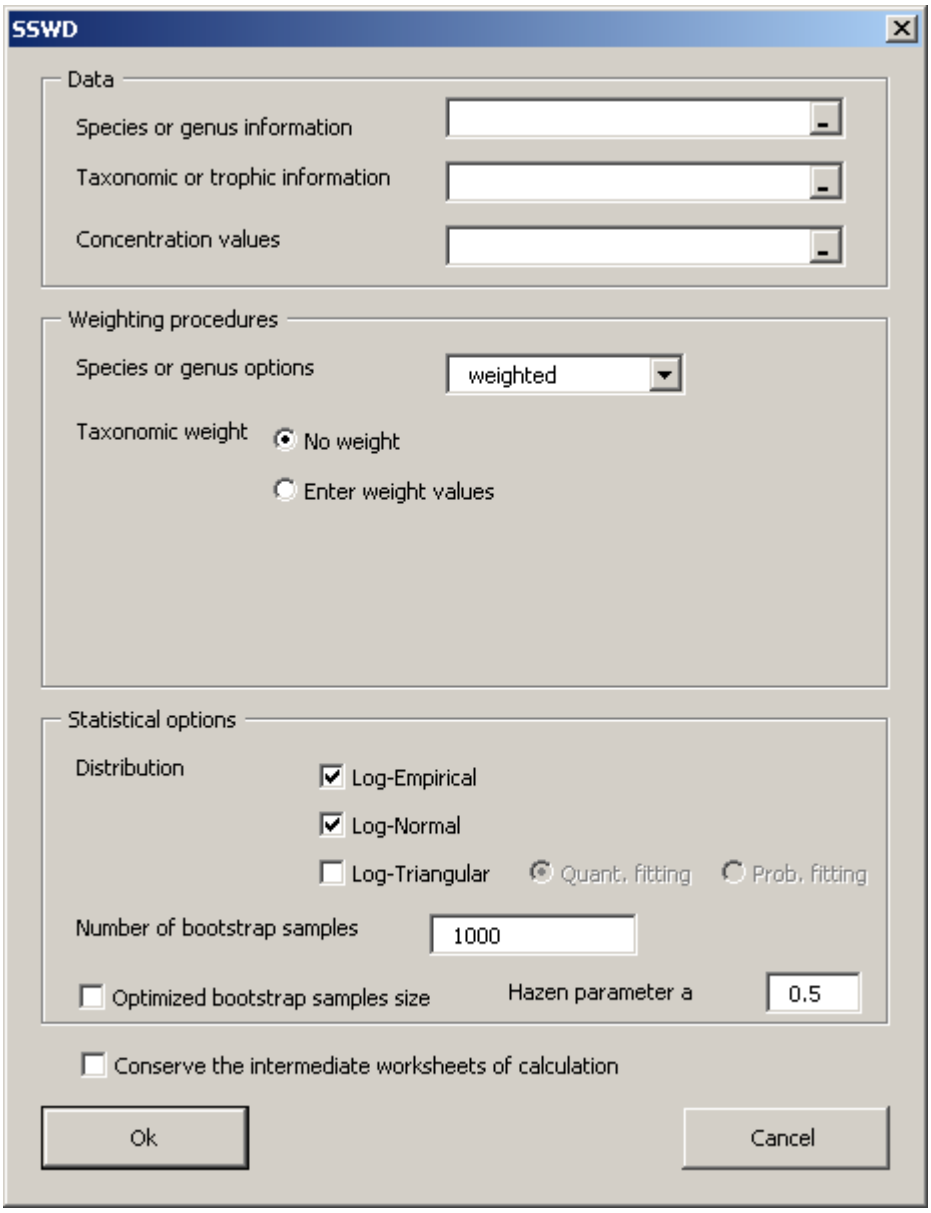
   It must also not contain any space. But it can contain the characters:

   . _



	1	2	3
1	**Species**	**Trophic level**	**Concentration in µg/l**
2	Microcystis aeruginosa	Algae	2.5
3	Daphnia magna	Invertebrates	15
4	Uronema parduczi	Invertebrates	21
5	Anabaena inequalis	Algae	25
6	Salmo gairdneri	Vertebrates	62
7	Entosiphon sulcatum	Invertebrates	70
8	Daphnia magna	Invertebrates	90
9	Salmo gairdneri	Vertebrates	104
10	Salmo gairdneri	Vertebrates	134
11	Salmo gairdneri	Vertebrates	238
12	Nostoc muscorum	Vertebrates	250
13	Pimephales promelas	Vertebrates	380
14	Pimephales promelas	Vertebrates	380
15	Chilomonas paramecium	Invertebrates	410
16	Salmo gairdneri	Vertebrates	431
17	Scenedesmus quadricauda	Algae	650
18	Salmo gairdneri	Vertebrates	700
19	Salmo gairdneri	Vertebrates	1100
20	Chlorella pyrenoidosa	Algae	4000
21	Chlamydomonas sp	Algae	10000
22			

3. From the main menu of Excel, select **SSWD ▶ Run** to start the SSWD tool. A dialog box must appear as follows:



Fill in the SSWD dialog box

Data

1. On the **Species or genus information** line, select the range of data or the column containing the name of the species of each ecotoxicological test or the genus of the species of each test, with the heading in the first line of the selection. If you select a column, the first line of the column must contain the heading and the program stops reading the data at the first empty cell.

2. On the **Taxonomic or trophic information** line, select the range of data or the column (heading included) containing the information about the taxonomic group or the trophic level.

3. On the **Concentration values** line, select the range of data or the column (heading included) containing the ecotoxicological test results.

Weighting procedures

1. Line **Species or genus options**, three options are proposed to account for redundant data for each species or genus:

   - **Weighted**: all the data are used, but each piece of data is weighted in order to give each species (or genus) the same weight within the SSWD. Intra-species variation is taken into account, but no species is given more importance than any other.

   - **Unweighted**: the available data are treated as if they all belonged to different species. Intra-species variation is taken into account, but there is a risk of giving too much importance to one species (or genus) if the number of the observed data for one species is greater than the number of data available for the others.

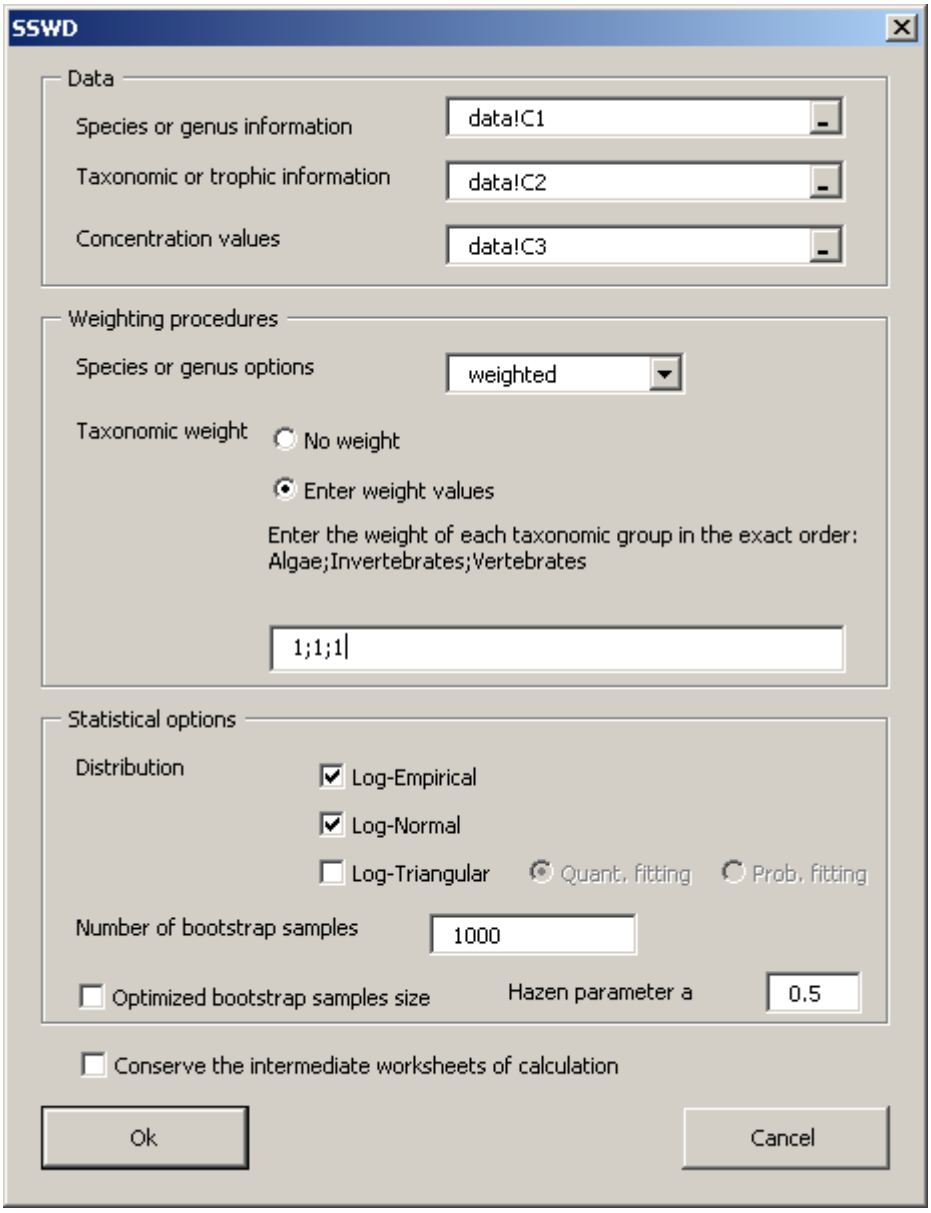
   - **Mean**: a geometric mean of data for each species (or genus) is first calculated; in this way a new set of data is obtained, in which for each species (or genus) there is one single corresponding data point. In this case, intra-species variations are ignored.

2. Line **Taxonomic weight**, two approaches are proposed regarding the proportions of data of each taxonomic group or each trophic level:

   - If you select **No Weight**, the available data are treated without caring about the proportions of data of each taxonomic group (or trophic level) in the database. But, these proportions may influence directly the result of the SSD, for example if one taxonomic group (or trophic level) is predominant in number.

   - With the option **Enter weight values**, you can allocate weight to each taxonomic group or trophic level. In particular, the data can be balanced, in such a way that the taxonomic groups are equally weighted in the SSWD. For this, put the same weight (1 for example) for each taxonomic group. Or else the data can be weighted so to respect different proportions in the taxonomic groups (or trophic level). For example, you can choose the proportions proposed by Forbes and Calow (10%, 26% and 64% respectively for vertebrates, invertebrates and algae [12]); for this, put the weight values 10, 26 and 64 (or 0.1; 0.26; 0.64) respectively for vertebrates, invertebrates and algae.

   If you select this last option, put the weight values in the alphabetical order of the taxonomic groups (or trophic levels); each weight must be separated by semi-colon. See below.



A weight is now allocated to each concentration data according to the species (or genus) options and to the taxonomic (or trophic) weight: the used data are transformed into weighted data points (see § Weighted data points and Weighted cumulative probability).

Statistical options

1. Select the **Distribution** to apply to the weighted data points. You can select the three distributions **log-empirical**, **log-normal** and **log-triangular** at the same time. Be careful with the empirical distribution, because, in this case, the best-estimate value and the confidence limits of the hazardous concentration are valid only if the number of data is important (see § Log-empirical distribution).

2. The parameters of the **log-triangular** distribution are obtained by fitting the theoretical distribution to the empirical weighted data points. Two possibilities are offered: fitting the quantiles (**Quant. fitting**); fitting the cumulative empirical

probabilities of the data (**Prob. fitting**). Test the two methods and retain the best fit (according to the multi-R-square coefficient for example, see § Log-triangular distribution).

3. Confidence limits of the hazardous concentration are estimated by bootstrap (see [3,9,17] and § Weighted bootstrapping). Enter the **Number of bootstrap samples** generated to calculate the confidence limits. Be careful with the log-triangular distribution, because it costs much greater computational time than the other distributions: test it before with a small **Number of bootstrap samples**.

4. The number of data points drawn in each sample of the bootstrap approach (the samples size) is by default equal to the number of used data (bootstrap *n* out of *n* [3,9]). If you select the option **Optimized bootstrap samples size**, the number of data points drawn in each sample of the bootstrap is optimized regarding the proportions of data and the weight of each taxonomic group or trophic level (see [17] and § Optimized bootstrap samples size).

   Note that the number of drawn data points (in the bootstrap samples) will always be less than 250, because of the columns number limits of Excel. If the number of ecotoxicological data used is greater than 250, the program selects a number of drawn data points equal to 250 by default.

5. The Hazen method enables to allocate a cumulative empirical probability to each data point of a sample. It is used to put the empirical data points in the graphical representation of the distribution, to estimate the log-empirical hazardous concentration as well as to estimate the log-triangular distribution parameters. The value of the **Hazen parameter *a*** is 0.5 by default. This value influences the values of the log-empirical and the log-triangular hazardous concentration and the SSWD graphics for the three distributions. Do not change this value without reading the paragraph Weighted cumulative probability.

Start calculation

Click on the **OK** button to start calculation. The results with the graphics are put in a new worksheet named **SSWD_result** (see § Results). If you want to restart the SSWD procedure and to conserve the first results, change the name of the **SSWD_result** worksheet.

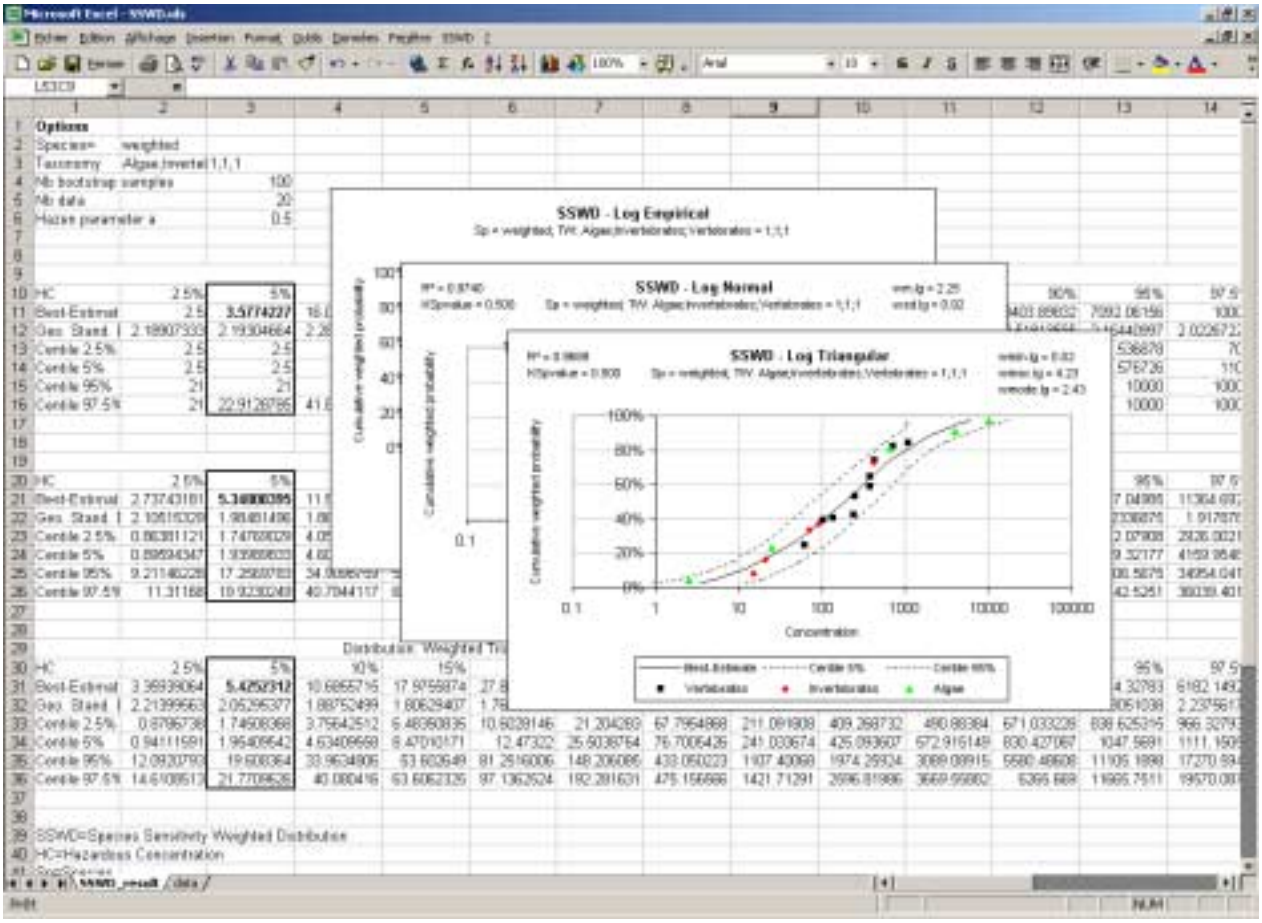
You can **Conserve the intermediate worksheets of calculation**, if you select this option. In this case, in addition to the **SSWD_result** worksheet, the program display the worksheet of:

- the weighted data points used to the SSWD best-estimate calculation (weight_result);
- the samples generated by the bootstrap procedure (draw_result);
- the intermediate results for the log-empirical distribution (qemp_result);
- the intermediate results for the log-normal distribution (qnorm_result);
- the bootstrap samples increasingly sorted (draw_sort);
- the intermediate results for the log-triangular distribution with the probability-fitting option (ftriang_result);
- the intermediate results for the log-triangular distribution with the quantiles-fitting option (qtriang_result).

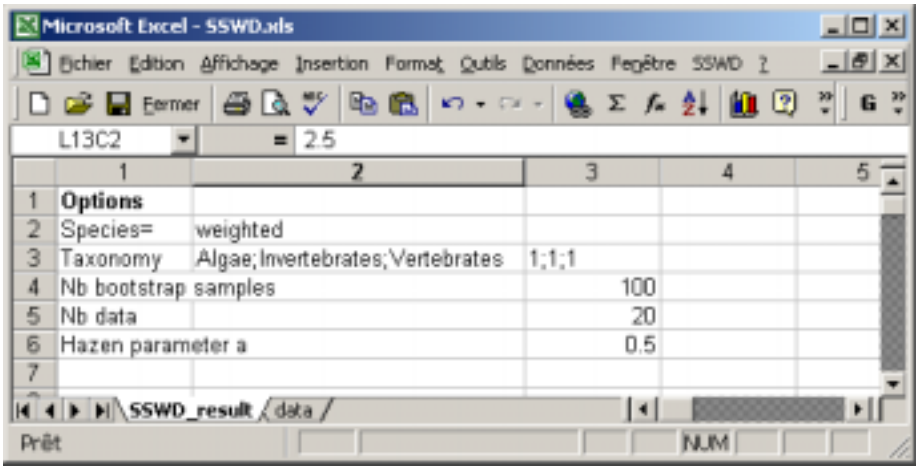
Displaying these worksheets is not usually necessary; all the important information is put in the **SSWD_result** worksheet.

Results

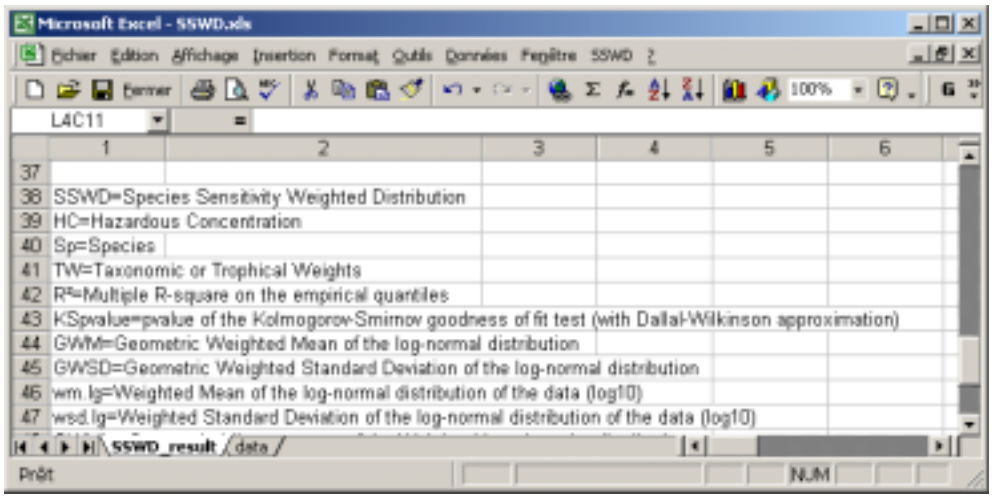
The numerical results are put in a worksheet named **SSWD_result** and the SSWD graphics are displayed directly in the same worksheet. See below.



The user finds in this worksheet information about the chosen options: the weight options respectively for species (or genus) and trophic levels (or taxonomic groups), the number of bootstrap samples, the number of data drawn in each bootstrap sample and the Hazen parameter a. See below.

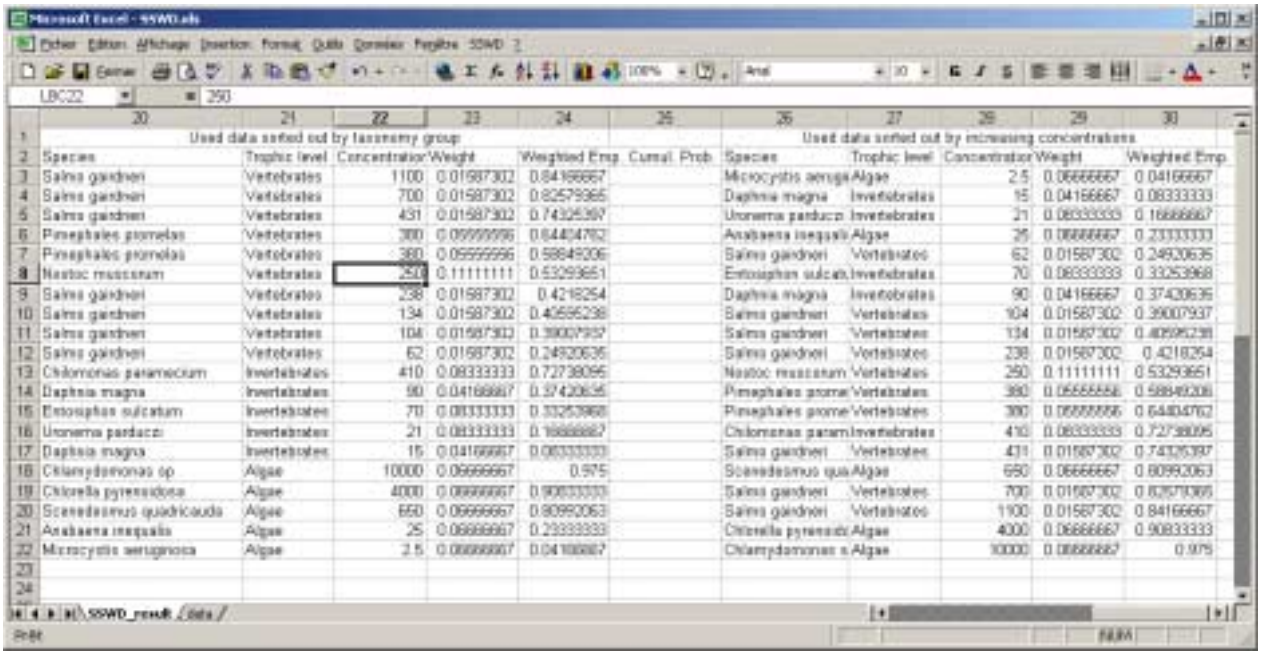


All the used acronyms are defined under the results in the same worksheet. See below.



Weighted data

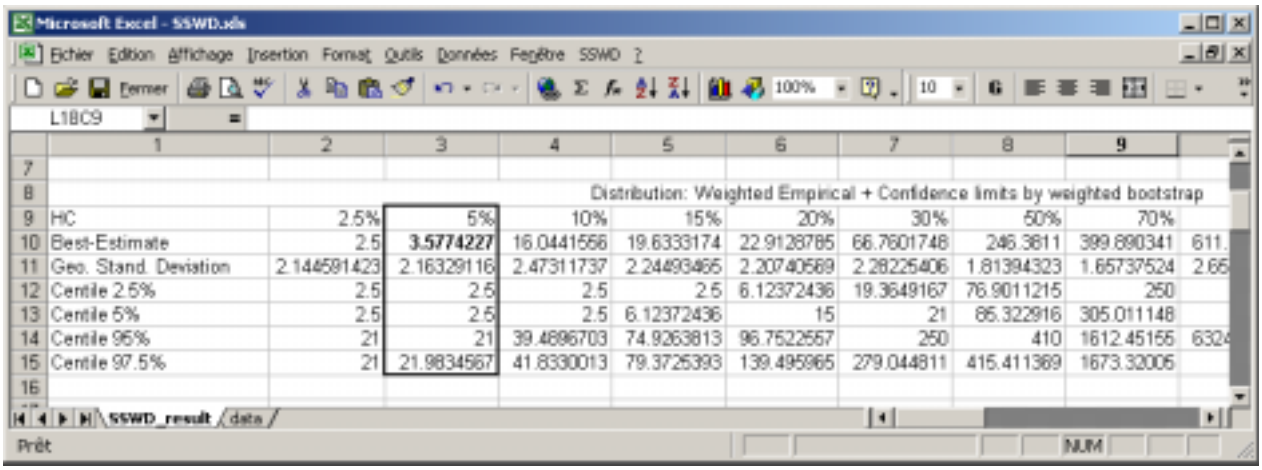
Data, weight (see § Weighted data points) and weighted cumulative probability (see § Weighted cumulative probability) of each data are put in two tables in the results worksheet on the right of the hazardous concentration values. In one table, data are sorted out by taxonomic group or trophic level. In the other one, data are sorted out by increasing concentration. These two tables are used by the SSWD graphics; don't remove it. See below.



Hazardous concentrations

For each chosen distribution, calculated hazardous concentrations are displayed in a table with the best-estimate values, the geometrical standard deviations of these

values, and their confidence limits at 90% and 95% (centiles 5 and 95%, 2.5 and 97.5% respectively). Different hazardous threshold are proposed in column: $HC_{2.5\%}$, $HC_{5\%}$, $HC_{10\%}$, etc. In particular, the $HC_{5\%}$ values are framed and the best-estimate value are in bold type. By definition, this concentration may affect 5% of the species; so it protects 95% of the species with 50% confidence. The centile 2.5% (for example) of this value protects 95% of the species with 97.5% confidence. See below.



The concentration values are directly put in the unity of the data supplied by the user. All the table is used in the SSWD graphic; don't remove it. See below.

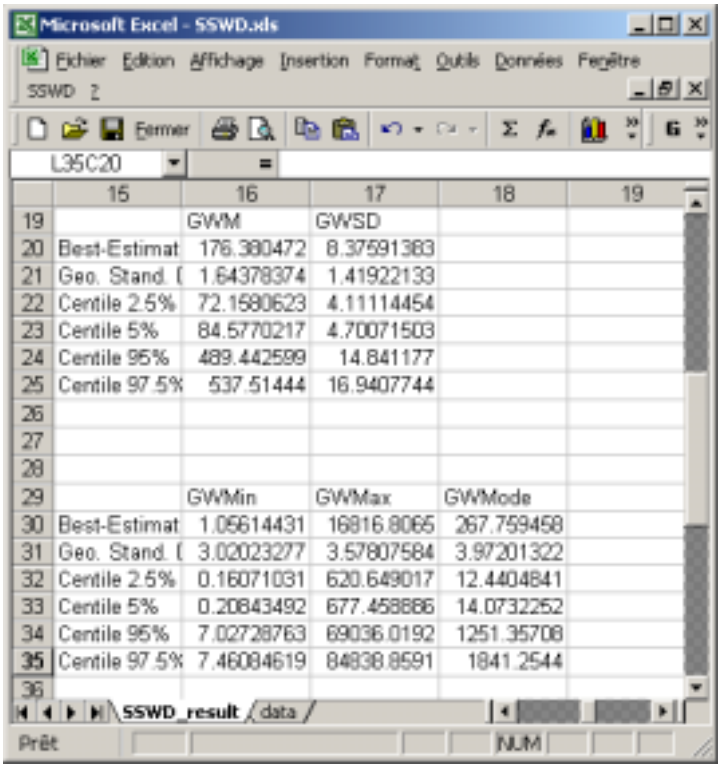
Note that the geometric standard deviation of a sample is the exponential of the standard deviation calculated on the log of the data of the sample.

Distribution parameters

For the theoretical distributions, the user find at the end of the hazardous concentration tables the best-estimate values of the distribution parameters, as well as the geometrical standard deviations, and the confidence limits at 90 and 95% of these estimations.

For the log-normal distribution, the displayed parameters are geometric weighted mean of the concentration data (*GWM*) and geometric weighted standard deviation of the data (*GWSD*, see § Log-normal distribution).

For the log-triangular distribution, the displayed parameters are geometric weighted min (*GWMin*), max (*GWMax*) and mode (*GWMode*, see § Log-triangular distribution). See below. Be careful, if the mode value is greater than the max value or lower than the min value, the estimation of the min, max and mode parameters is not good. Try an another estimation method (fitting the quantiles or the probabilities) or an another distribution.



Microsoft Excel - SSWD.xls

Fichier Edition Affichage Insertion Format Outils Données Fenêtre

SSWD 2

Fermer

L35C20 =

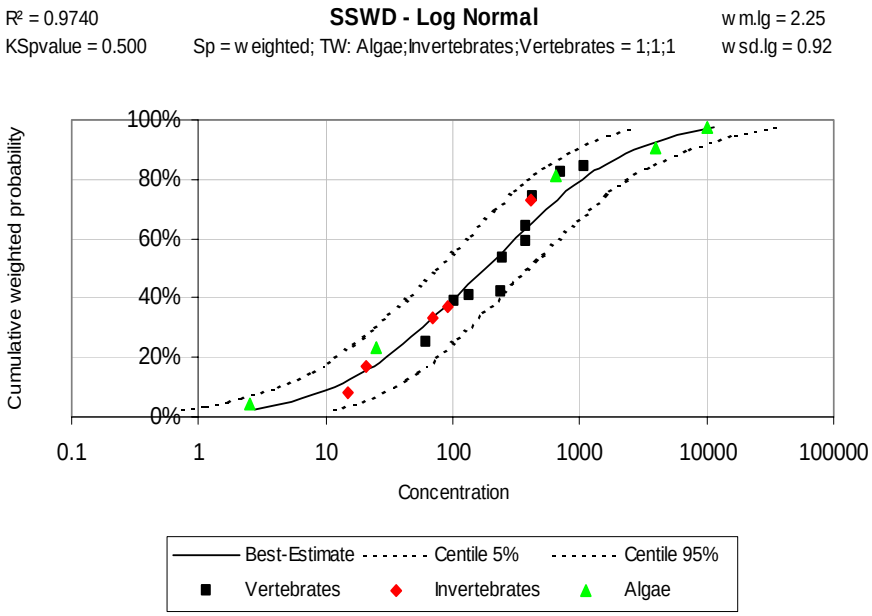
	15	16	17	18	19
19		GWM	GWSD		
20	Best-Estimat	176.380472	8.37591383		
21	Geo. Stand. [	1.64378374	1.41922133		
22	Centile 2.5%	72.1580623	4.11114454		
23	Centile 5%	84.5770217	4.70071503		
24	Centile 95%	489.442599	14.841177		
25	Centile 97.5%	537.51444	16.9407744		
26					
27					
28					
29		GWMin	GWMax	GWMode	
30	Best-Estimat	1.05614431	16816.8065	267.759458	
31	Geo. Stand. [	3.02023277	3.57807584	3.97201322	
32	Centile 2.5%	0.16071031	620.649017	12.4404841	
33	Centile 5%	0.20843492	677.458886	14.0732252	
34	Centile 95%	7.02728763	69036.0192	1251.35708	
35	Centile 97.5%	7.46084619	84838.8591	1841.2544	
36					

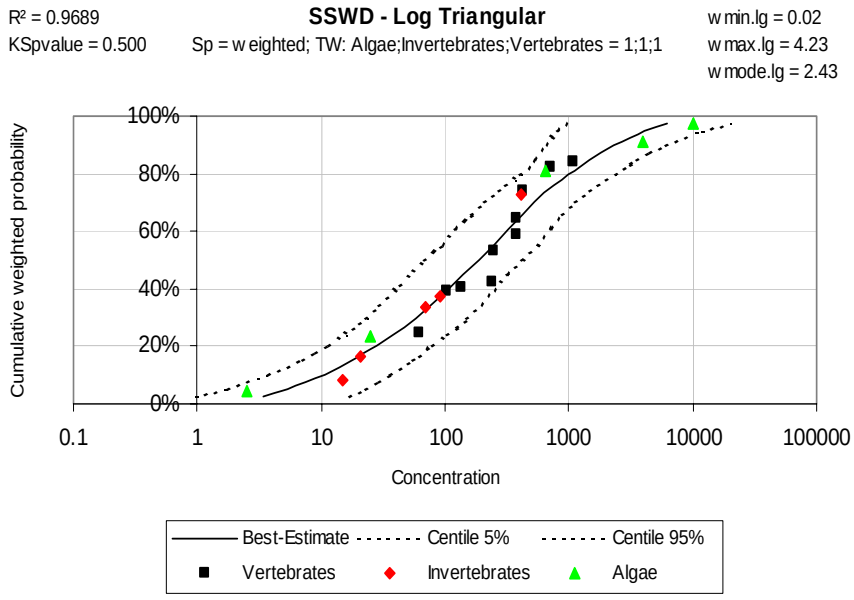
SSWD_result / data /

Prêt NUM

SSWD graphics

On the SSWD graphics are displayed the weighted data points, the empirical or the theoretical distribution and the 90% confidence limits of this distribution. The color of the data points depends on the taxonomic group or the trophic level of the considered species. For the theoretical distribution, are displayed the multiple R-square coefficient (R^2) between theoretical and empirical distribution and the Kolmogorov-Smirnov goodness of fit test (*KSpvalue*, see § Goodness of fit test). In addition, the user finds the parameters of the distribution calculated on the log of the data. For the log-normal distribution, it is the mean (*wm.lg*) and the standard deviation (*wsd.lg*, see § Log-normal distribution). For the log-triangular distribution, it is the min, the max and the mode (*wmin.lg*, *wmax.lg* and *wmode.lg*, see § Log-triangular distribution). See below.





In these graphics, the concentration values are directly put in the unity of the data supplied by the user, but with a log-scale. These graphics are constructed with the worksheet tables and can be modified by the user. In particular, the 90% confidence limits can be replaced by the 95% confidence limits.

METHODOLOGY

Weighting procedure

Weighted data points

According to the options about species (or genus) and taxonomy (or trophic level), a weighting coefficient is allocated to each concentration data. This weight is mathematically defined below. We note w_{ijk}^{d} the weighting coefficient of the data *k* of the species (or genus) *j* to the taxonomic group or trophic level *i* (the little *d* mean that it is the weight of each data and not the weight of the taxonomic group or trophic level). See it in the table.

Taxonomy/Species options	Unweighted	Mean	Weighted
No weight	$\frac{1}{n}$	$\frac{1}{ns}$	$\frac{1}{n_{ij} \times ns}$
Weight w_i for each taxonomic group *i* Standardized weight: $w_i^{\otimes} = w_i \Big/ \sum_{j=1}^{nt} w_j$	$\frac{w_i^{\otimes}}{n_{i.}}$	$\frac{w_i^{\otimes}}{ns_i}$	$\frac{w_i^{\otimes}}{n_{ij} \times ns_i}$

n_{ij} is the number of data associated with species *j* in taxonomic group *i*; $n_{i.}$ is the number of data in taxonomic group *i*; *n* is the total number of data; ns_i is the number of different species in taxonomic group *i* and *ns* is the total number of different species; *nt* is the number of taxonomic groups (or trophic levels). All the data belonging to the same species have the same weight.

The weighting coefficients are applied to the whole dataset for the **Unweighted** and **Weighted** species options, but to the species (or genus) mean values for the **Mean** option. The number of data in this latter case corresponds to the number of different species represented.

Finally, note that the weighting coefficients are automatically standardized:

$$\sum_{i,j,k} w_{ijk}^{d} = 1 .$$

Weighted cumulative probability

The empirical cumulative probability of each data of a sample is a ambiguous concept; several estimations are possible. The classical method consists of sorting increasingly the data and allocating the probability i/n to each one, where *n* is the total number of data and *i* the rank of each one. Nevertheless, this approach is not symmetrical because the smallest data has the probability $1/n$ and the greatest the probability 1. So we prefer the Hazen method.

The Hazen method [11] enables to allocate a cumulative empirical probability to each data point of a sample in a symmetrical manner. By this method, if i is the rank of each data point x_i in the sample and n the total number of data, the cumulative empirical probability p_i of x_i is defined as:

$$p_i = \frac{i-a}{n+1-2\times a} \quad \text{with generally } a = 0.5 .$$

The Hazen method was modified to estimate the empirical "weighted" cumulative probability of each data point. The modified Hazen method is mathematically defined as follows:

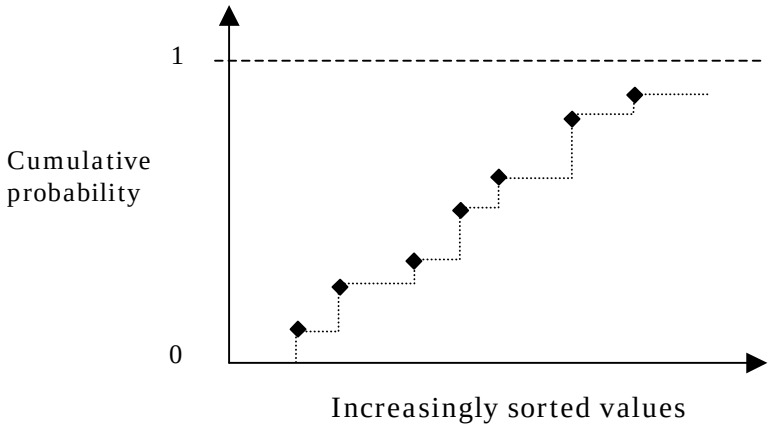
If x_i, i=1 to n are the increasingly sorted data points and w_i^d the weighting coefficient previously defined of each data, let's define $r_i = n \times \sum_{j=1}^{i} w_j^d$ the weighted-rank of each data point x_i.
The weighted cumulative probability p_i^w of each data point x_i will be defined as:

$$p_i^w = \frac{r_i - a}{n+1-2\times a} \quad \text{with } a = 0.5 \text{ by default} .$$

Unfortunately, if the first data points have too very small weights, their weighted-rank values r_i may be less than a, and the calculation of p_i^w with $a = 0.5$ is not possible. So, in this case, the program selects automatically $a = 0$ and:

$$p_i^w = \frac{r_i}{n+1} .$$

The weighted cumulative probabilities are used for the graphics of the species sensitivity weighted distribution: it enables to place the weighted data points in a PDF (probability distribution function) graph, whatever the chosen theoretical distribution. See below.



The weighted cumulative probabilities are used as well to estimate the log-empirical hazardous concentration and the log-triangular distribution parameters (see § Log-empirical distribution and Log-triangular distribution).

Calculation of the best-estimate value of the hazardous concentration

Log-empirical distribution

With the log-empirical distribution, the calculation of the $HC_{th\%}$ is defined by linear interpolation on the log of the data as follows: if x_i, i=1 to n are sorted increasingly data points and p_i^w the weighted cumulative probability of each one, we have:

$$\begin{cases} HC_{th} = x_1 \quad \text{if } th \le p_1^w \\ HC_{th} = \exp\left\{\log(x_{i+1}) - \dfrac{[\log(x_{i+1}) - \log(x_i)] \times (p_{i+1}^w - th)}{(p_{i+1}^w - p_i^w)}\right\} \quad \text{if } p_i^w \le th < p_{i+1}^w \\ HC_{th} = x_n \quad \text{if } th \ge p_n^w \end{cases}$$

Be careful, estimation of the $HC_{th\%}$ when th is lower than the lowest cumulative empirical probability of the data is not correct, because, in this case, the $HC_{th\%}$ value is limited by the smallest value (data) of the sample.

Log-normal distribution

To fit the log-normal distribution to the weighted concentration data, the weighted mean μ and standard deviation σ are calculated on the log data as follows:

$$\begin{cases} \mu = \sum_{i=1}^{n} \log(x_i) \times w_i^d \\ \sigma = \sqrt{\dfrac{n}{n-1} \sum_{i=1}^{n} [\log(x_i) - \mu]^2 \times w_i^d} \end{cases}$$

where: n is the number of used data; w_i^d is the weighting coefficient associated with data point x_i.

Comments: a ratio $n/(n-1)$ is introduced into the weighted standard deviation; in this way an unbiased estimator of the population standard deviation is obtained. If the weighting coefficient w_i^d equals $1/n$ (option species=**Unweighted** and taxonomy=**No weight**), the preceding formulas correspond exactly to the classic one for the mean and for standard deviation.

With the log-normal distribution, the calculation of the $HC_{th\%}$ is defined as follows:

$$HC_{th} = \exp[\text{NormalQuantile}(th\%, \mu, \sigma)]$$

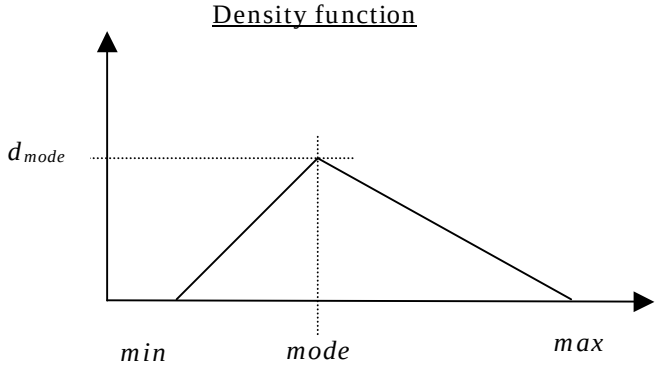
(We use the NormalQuantile function of Excel).

Note that we call geometric mean and geometric standard deviation respectively the exponential of the mean calculated on the log-data and the exponential of the standard deviation calculated on the log-data.

Log-triangular distribution

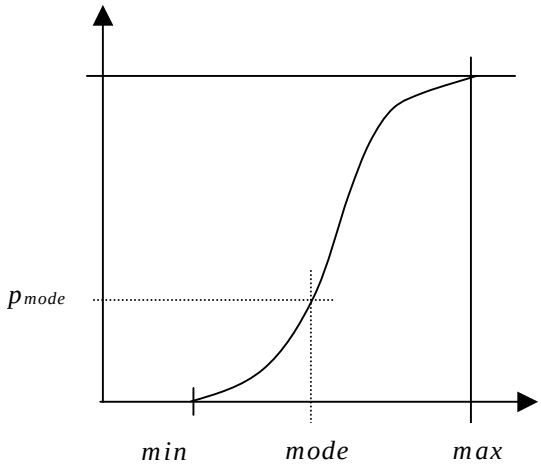
The triangular distribution is defined by three parameters: *min*, *max* and *mode*. See below:

Density function



With $d_{mode} = \dfrac{2}{max - min}$

Probability distribution function



With $p_{mode} = \dfrac{mode - min}{max - min}$

The density function of the triangular distribution is defined as follows:

$$f(x) = \begin{cases} 0 \quad \text{pour } x < min \\ \dfrac{2 \times (x - min)}{(max - min)(mode - min)} & \text{pour } min \le x \le mode \\ \dfrac{-2 \times (x - max)}{(max - min)(max - mode)} & \text{pour } mode \le x \le max \\ 0 \quad \text{pour } x > max \end{cases}$$

The parameters of the log-triangular distribution are estimated by fitting the theoretical distribution to the weighted data points. Two approaches are possible: (1) fitting the quantiles; (2) fitting the cumulative probabilities of the data.

(1) Fitting the quantiles

The theoretical quantiles of the triangular distribution with parameters *min*, *max* and *mode* corresponding to cumulative probability *p* is defined below:

$$\mathrm{Q_t}(p, min, max, mode) = \begin{cases} min + \sqrt{p \times (max - min) \times (mode - min)} & \text{if } p \le p_{mode} \\ max - \sqrt{(1-p) \times (max - min) \times (max - mode)} & \text{if } p \ge p_{mode} \\ \text{with } p_{mode} = \dfrac{mode - min}{max - min} \end{cases}$$

So to find the best-estimate value of the *min*, *max* and *mode* parameters, the log of the triangular quantile function (previously defined) is fitted to the log data points by minimizing the sum of square of the residuals (*SSRq*) defined as follows:

$$SSRq = \sum_{i=1}^{n} \left\{ \log\left[\mathrm{Q_t}\left(p_i^w, min, max, mode\right)\right] - \log(x_i) \right\}^2$$

where x_i (i=1, 2,… n) are the concentration data and p_i^w the weighted cumulative probability of each one. For this minimization, the complementary macro solver of Excel is used.

The best-estimate hazardous concentration $HC_{th\%}$ is then calculated as follows:

$$HC_{th} = \exp\left[\mathrm{Q_t}\left(th, \overline{min}, \overline{max}, \overline{mode}\right)\right]$$

where $\overline{min}$, $\overline{max}$ and $\overline{mode}$ are the values previously estimated.

(2) Fitting the probabilities

The theoretical cumulative probability of the triangular distribution with parameters *min*, *max* and *mode* corresponding to the quantile *q* is defined below:

$$\mathrm{P_t}(q, min, max, mode) = \begin{cases} 0 \text{ if } q \le min \\ \dfrac{(q - min)^2}{(max - min) \times (mode - min)} \text{ if } min \le q \le mode \\ 1 - \dfrac{(q - max)^2}{(max - min) \times (max - mode)} \text{ if } mode \le q \le max \\ 1 \text{ if } q \ge max \end{cases}$$

So to find the best-estimate value of the *min*, *max* and *mode* parameters, the triangular cumulative probability function (previously defined) is fitted to the empirical weighted cumulative probability of each data point by minimizing the sum of square of the residuals (*SSRq*) defined as follows:

$$SSRp = \sum_{i=1}^{n} \left[\mathrm{P_t}\left(\log(x_i), min, max, mode\right) - p_i^w\right]^2$$

where x_i (i=1, 2,… n) are the concentration data and p_i^w the weighted cumulative probability of each one. For this minimization, the complementary macro solver of Excel is used.

The best-estimate hazardous concentration $HC_{th\%}$ is then calculated as follows:

$$HC_{th} = \exp\left[\mathrm{Q}_t\left(th, \overline{min}, \overline{max}, \overline{mode}\right)\right]$$

where $\overline{min}$, $\overline{max}$ and $\overline{mode}$ are the values previously estimated.

Note that we call geometric min, max and mode, respectively the exponential of the min, max and mode parameters calculated on the log of the data.

Be careful, if the mode value is greater than the max value or lower than the min value, the estimation of the min, max and mode parameters is not good. Try an another estimation method (fitting the quantiles or the probabilities) or an another distribution.

Goodness of fit test

The goodness of fit (for the log-normal and the log-triangular distributions) is tested by a Kolmogorov-Smirnov (*KS*) test with Dallal-Wilkinson approximation.

The *KS* statistic is defined as $KS = \sup\left(\left|p_i^w - p_i\right|\right)$,

where p_i^w is the weighted empirical cumulative probability of each sorted data point x_i and p_i the cumulative probability corresponding to the parametric distribution.

For the log-normal distribution, p_i is defined as:

$$p_i = \mathrm{NormalCumulativeProbabiliy}\left(\log\left(x_i\right), \mu, \sigma\right)$$

Where μ and σ are respectively the weighted mean and standard deviation estimating with the log of the data points x_i. (We use the NormalCumulativeProbability function of Excel).

For the log-triangular distribution, p_i is defined as follows:

$$p_i = \begin{cases} 0 \text{ if } \log\left(x_i\right) \le min \\ \dfrac{\left(\log\left(x_i\right) - min\right)^2}{\left(max - min\right) \times \left(mode - min\right)} \text{ if } min \le \log\left(x_i\right) \le mode \\ 1 - \dfrac{\left(\log\left(x_i\right) - max\right)^2}{\left(max - min\right) \times \left(max - mode\right)} \text{ if } mode \le \log\left(x_i\right) \le max \\ 1 \text{ if } \log\left(x_i\right) \ge max \end{cases}$$

Where *min*, *max* and *mode* are the parameters of the triangular distribution fitted to the weighted log data points.

The target of goodness of fit test is to estimate the probability (p_{value}) of making a mistake when we say that the chosen parametrical distribution is not the good one, that is the empirical distribution is not log-normal or log-triangular. This p_{value}

corresponding to the *KS* statistic by the Dallal-Wilkinson approximation is described below:

We note *n* the number of used data points.
If *n*<5 the p_{value} cannot be estimated, so $p_{value} = 0$ by default.
Else:

$$p_{value} = \exp\begin{pmatrix} -7.01256 \times KS^2 \times (n + 2.78019) + 2.99587 \times KS \times \sqrt{n + 2.78019} \\ -0.122119 + 0.974598/\sqrt{n} + 1.67997/n \end{pmatrix}$$

If *n*>100, we take $KS = KS \times (n/100)^{0.49}$ and $n = 100$ in the rest of the formula.

If $p_{value} > 0.1$ then $p_{value} = 0.5$ because the Dallal-Wilkinson approximation is most accurate for $pvalue \leq 0.1$. The lowest the *pvalue*, the lowest the confidence in the chosen distribution.

In addition to the Kolmogorov-Smirnov goodness of fit test, the multiple R-square coefficient between empirical and theoretical distribution is calculated. This coefficient is defined on the quantiles as follows:

If x_i (with *i*=1 to *n*) are the sorted increasingly data point and p_i^w the weighted empirical cumulative probability of each one, we note q_i the theoretical quantile corresponding to the p_i^w probability.

For the log-normal distribution, q_i is defined as:

$$q_i = \text{NormalQuantile}(p_i^w, \mu, \sigma)$$

Where μ and σ are respectively the weighted mean and standard deviation estimating with the log of the data points x_i. (We use the NormalQuantile function of Excel).

For the log-triangular distribution, q_i is defined as follows:

$$q_i = \begin{cases} min + \sqrt{p_i^w \times (max - min) \times (mode - min)} \text{ if } p_i^w \leq p_{mode} \\ max - \sqrt{(1 - p_i^w) \times (max - min) \times (max - mode)} \text{ if } p_i^w \geq p_{mode} \\ \text{with } p_{mode} = \dfrac{mode - min}{max - min} \end{cases}$$

Where *min*, *max* and *mode* are the parameters of the triangular distribution fitted to the weighted log data points.

So if we define $e_i = q_i - \log(x_i)$ the deviation between the data points and the corresponding theoretical quantiles (i.e. the residuals), the multiple R-square coefficient is calculated as follows:

$$R^2 = 1 - \frac{\mathrm{var}_w(e_i)}{\mathrm{var}_w[\log(x_i)]}$$

where $\mathrm{var}_w(e_i)$ and $\mathrm{var}_w[\log(x_i)]$ are respectively the weighted variance of the residuals and the one of the log-data. These parameters as defined below:

$$\begin{cases} \mathrm{var}_w[\log(x_i)] = \dfrac{n}{n-1}\displaystyle\sum_{i=1}^{n}[\log(x_i) - \mu]^2 \times w_i^d \quad \text{with } \mu = \sum_{i=1}^{n}\log(x_i) \times w_i^d \\ \mathrm{var}_w(e_i) = \dfrac{n}{n-1}\displaystyle\sum_{i=1}^{n}(e_i - \mu_e)^2 \times w_i^d \quad \text{with } \mu_e = \sum_{i=1}^{n} e_i \times w_i^d \end{cases}$$

where n is the number of used data; w_i^d is the weighting coefficient associated with the data point x_i. The multiple R-square is necessary lower than 1. The theoretical distribution with the best value of the multiple R-square must be chosen.

Estimating the confidence limits of the HC

Weighted bootstrapping

Bootstrap is a data resampling approach used to estimate confidence limits of different statistics (see [9]). It was used in the case of the species sensitivity distribution by Verdonck [10], Wheeler [13] and Grist [15], for examples. The weighted bootstrap [17] is defined as a non-equiprobable resampling of data with replacement: the probability of drawing each data point corresponds to the weighting coefficient previously defined. The number of data drawn for each sample of the bootstrap corresponds to that of the initial dataset (bootstrap *n* out of *n*), or to an optimized *m* value (see § Optimized bootstrap samples size). The number of bootstrap samples is defined by the user; this *B* value must be greater than 1000 to have a correct estimation of the confidence limits. For each sample, a $HC^{*}_{th\%}$ is calculated with the selected distribution (log-empirical, log-normal or log-triangular); so *B* values of $HC^{*}_{th\%}$ are calculated by the bootstrap approach. With these values, an estimation of the confidence limits of the best-estimate hazardous concentration $HC_{th\%}$ is given, in particular, the percentile 2.5%, 5%, 95% and 97.5%. We note with a "*" the values of the parameters estimated on a bootstrap sample. A bias-correction by the best-estimate value is applied for the log-normal and log-triangular distributions. The bootstrap approach can be used if the number of available data is greater than 10-15 for the log-normal or log-triangular distribution; but for the log-empirical distribution, the number of data must be greater (see § Log-empirical distribution).

The bootstrap procedure was developed with the multinomial random generator of Excel.

Optimized bootstrap samples size

When chosen taxonomic (or trophic level) weights are significantly different than proportions of available data, some data may have a very high weight within the SSWD. These data will tend to be in redundancy in samples generated by the bootstrap: each sample contains the same value several times, yielding to an unsatisfactory distribution. The option **Optimized bootstrap samples size** compensates for this drawback.

Mathematically, this approach is expressed as follows: if the n_i (i=1,2,3,…) are the numbers of data observed in each taxonomic group (or trohpic level) and w_i the weight desired for each one, $w_i^{\otimes}$, the standardized weight, is defined by: $w_i^{\otimes} = w_i \Big/ \sum w_j$, then $m = \min(n_i / w_i^{\otimes}, i = 1,2,3,...)$ corresponds to the optimized number's draws for each samples of the bootstrap procedure (see [17]). In fact, the nearest whole round number to this value is used. This *m* value is necessary less than the number of available data (bootstrap *m* out of *n* with $m < n$). This approach cannot be used if the number of available data is too initially low. You must verify that the number of data drawn in each sample of the bootstrap by this procedure is greater than 10-15.

Log-empirical distribution

For each sample drawn by the bootstrap, the cumulative empirical probability of each data point is calculated by the direct Hazen method: if i is the rank of each (increasingly sorted) data point x_i^* in the bootstrap sample and n the total number of draws, the cumulative probability p_i^* of x_i^* is:

$$p_i^* = \frac{i-a}{n+1-2\times a} \quad \text{with generally } a = 0.5$$

So the calculation of the $HC_{th\%}^*$ of the bootstrap sample is defined by linear interpolation as follows:

$$HC_{th}^* = \begin{cases} x_1^* \text{ if } th \le p_1^* \\ \exp\left\{\log\left(x_{i+1}^*\right) - \dfrac{\left[\log\left(x_{i+1}^*\right) - \log\left(x_i^*\right)\right]\times\left(p_{i+1}^* - th\right)}{\left(p_{i+1}^* - p_i^*\right)}\right\} \text{ if } p_i^* \le th < p_{i+1}^* \\ x_n^* \text{ if } th \ge p_n^* \end{cases}$$

With B bootstrap samples, B values of $HC_{th\%}^*$ are calculated. The confidence limits of the HC_{th} best-estimate value are estimated with this B values sample. The 90% confidence limits of the best-estimate HC_{th} is defined by the values:

$$\exp\left[\text{Centile}\left(\log\left(HC_{th}^*\right), p = 0.05\right)\right] \text{and } \exp\left[\text{Centile}\left(\log\left(HC_{th}^*\right), p = 0.95\right)\right].$$

The 95% confidence limits of the best-estimate HC_{th} is defined by the values:

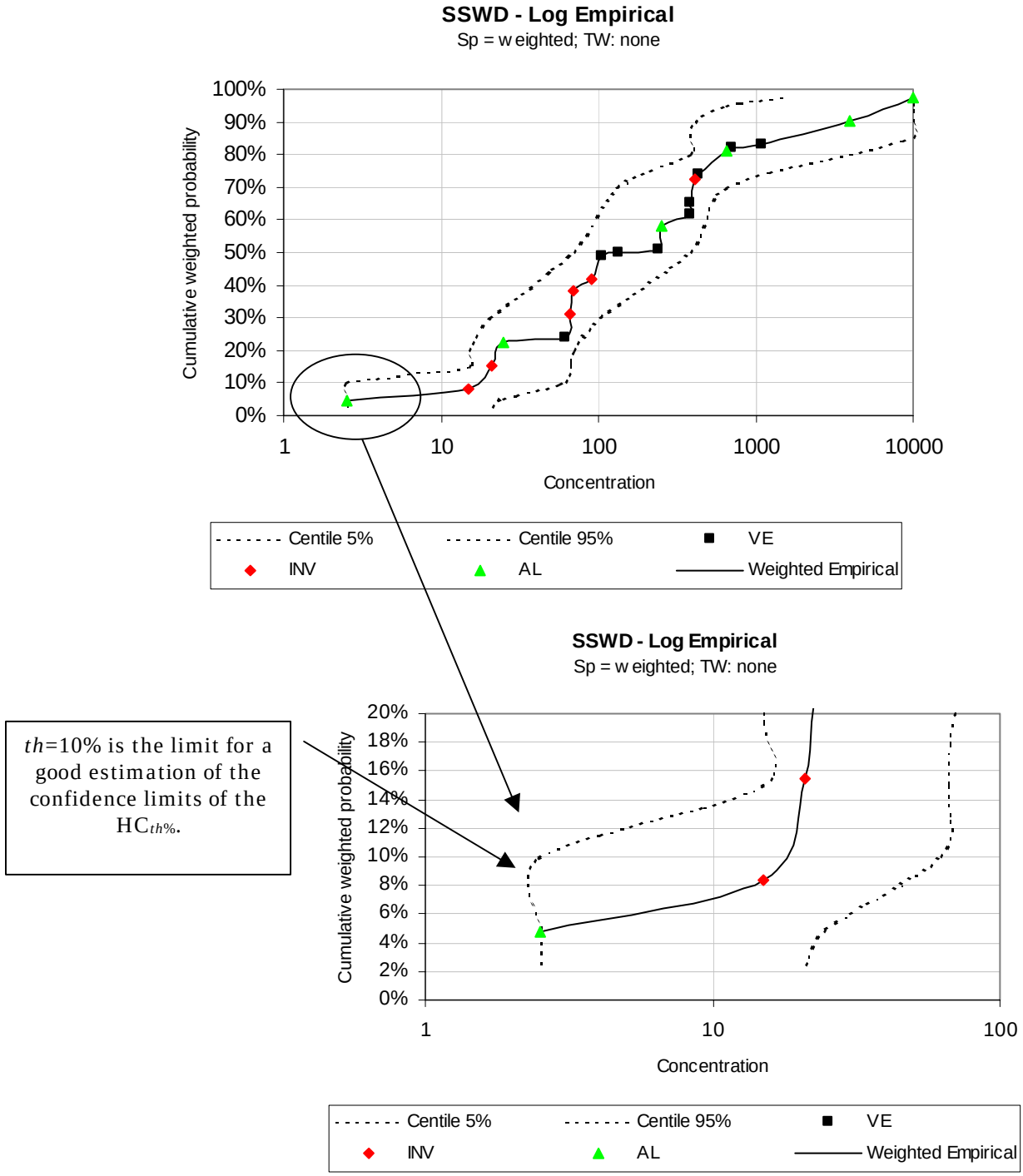
$$\exp\left[\text{Centile}\left(\log\left(HC_{th}^*\right), p = 0.025\right)\right] \text{and } \exp\left[\text{Centile}\left(\log\left(HC_{th}^*\right), p = 0.975\right)\right].$$

Where Centile is the empirical quantile function of Excel. This one is defined as follows:

If x_i, i=1 to n are increasingly sorted values of a sample X,

$$\text{Centile}\left(X, p\right) = \begin{cases} x_1 \text{ if } p = 0 \\ x_i - \left(x_i - x_{i-1}\right)\left[i - 1 - p\left(n-1\right)\right] \text{ if } \dfrac{i-2}{n-1} < p \le \dfrac{i-1}{n-1} \end{cases} \quad \text{with } 2 \le i \le n.$$

Be careful with the log-empirical distribution, estimating the confidence limits of the $HC_{th\%}$ by bootstrap with a low value of th (th=5% for example) is only correct if the number of data is important. We can see graphically if the number of data is sufficient or not: see below.



Log-normal distribution

The mean and the standard deviation (in log data) of each sample drawn by the bootstrap procedure are calculated as follows:

$$\begin{cases} \mu^* = \dfrac{1}{n}\sum_{i=1}^{n} \log\left(x_i^*\right) \\ \sigma^* = \sqrt{\dfrac{n}{n-1}\sum_{i=1}^{n}\left[\log\left(x_i^*\right) - \mu^*\right]^2} \end{cases}$$

Where x_i^* (i=1, 2 to n) are the data drawn in a bootstrap sample and n the size of the sample.

The calculation of the $HC_{th\%}^*$ of the bootstrap sample is defined by:

$$HC_{th}^* = \exp\left[\text{NormalQuantile}\left(th\%, \mu^*, \sigma^*\right)\right].$$

With B bootstrap samples, B values of $HC_{th\%}^*$ are calculated. The confidence limits of the HC_{th} best-estimate value are estimated with this B values sample. The 90% confidence limits of the best-estimate HC_{th} is defined by the values:

$$\exp\left[\text{Centile}\left(\log\left(HC_{th}^*\right), p\right) - \text{Centile}\left(\log\left(HC_{th}^*\right), 0.5\right) + \log\left(HC_{th}^{best-estimate}\right)\right]$$

$$\text{with } p = 0.05 \text{ and } p = 0.95.$$

and the 95% confidence limits with $p = 0.025$ and $p = 0.975$.

Where Centile is the empirical quantile function of Excel. This one is defined as follows:

If x_i, i=1 to n are increasingly sorted values of a sample X,

$$\text{Centile}\left(X, p\right) = \begin{cases} x_1 \text{ if } p = 0 \\ x_i - \left(x_i - x_{i-1}\right)\left[i - 1 - p\left(n-1\right)\right] \quad \text{if } \dfrac{i-2}{n-1} < p \le \dfrac{i-1}{n-1} \quad \text{with } 2 \le i \le n \end{cases}$$

Note that a bias-correction is applied to estimate the confidence limits with the log-normal distribution.

Log-triangular distribution

The parameters of the log-triangular distribution corresponding to the values of each bootstrap sample are estimated by fitting the theoretical distribution to the data. Like for the best-estimate value, two procedures are possible: (1) fitting the quantiles; (2) fitting the cumulative probabilities.

For each sample drawn by the bootstrap approach, the cumulative empirical probability of each data point is calculated by the direct Hazen method: if i is the rank of each data point x_i^* in the bootstrap sample and n the total number of draws (the sample size), the cumulative probability p_i^* of x_i^* is:

$$p_i^* = \frac{i-a}{n+1-2\times a} \quad \text{with } a = 0.5 \text{ by default .}$$

(1) Fitting the quantiles

The values min^*, max^*, and $mode^*$ of the triangular distribution fitted to each drawn sample are estimated by minimizing the sum of square of the residuals $SSRq$ (see below):

$$SSRq = \sum_i \left\{\log\left[\mathrm{Q_t}\left(p_i^*, min^*, max^*, mode^*\right)\right] - \log\left(x_i^*\right)\right\}^2$$

where x_i^* (i=1, 2 to n) are the data drawn in a bootstrap sample and p_i^* the cumulative probability of each one. $\mathrm{Q_t}$ is the quantile function for the triangular distribution (see § Calculation of the best-estimate value of the hazardous concentration: Log-triangular distribution).

(2) Fitting the probabilities

The values min^*, max^*, and $mode^*$ of the triangular distribution fitted to each drawn sample are estimated by minimizing the sum of square of the residuals $SSRp$ (see below):

$$SSRp = \sum_i \left[\mathrm{P_t}\left(x_i^*, min^*, max^*, mode^*\right) - p_i^*\right]^2$$

where x_i^* (i=1, 2 to n) are the data drawn in a bootstrap sample and p_i^* the cumulative probability of each one. $\mathrm{P_t}$ is the cumulative probability function for the triangular distribution (see § Calculation of the best-estimate hazardous concentration: Log-triangular distribution).

So the calculation of the $HC_{th\%}^*$ of the bootstrap sample is calculated as:

$$HC_{th}^* = \exp\left[\mathrm{Q_t}\left(th, \overline{min}^*, \overline{max}^*, \overline{mode}^*\right)\right]$$

where $\overline{min}^*$, $\overline{max}^*$ and $\overline{mode}^*$ are the values previously estimated by fitting the quantiles or the probabilities.

With B bootstrap samples, B values of $HC_{th\%}^*$ are calculated. The confidence limits of the HC_{th} best-estimate value are estimated with this B values sample. The 90% confidence limits of the best-estimate HC_{th} is defined by the values:

$$\exp\left[\mathrm{Centile}\left(\log\left(HC_{th}^*\right), p\right) - \mathrm{Centile}\left(\log\left(HC_{th}^*\right), 0.5\right) + \log\left(HC_{th}^{best-estimate}\right)\right]$$

$$\text{with } p = 0.05 \text{ and } p = 0.95 .$$

and the 95% confidence limits with $p = 0.025$ and $p = 0.975$.

Where Centile is the empirical quantile function of Excel. This one is defined as follows:

If x_i, i=1 to n are increasingly sorted values of a sample X,

$$\text{Centile}(X, p) = \begin{cases} x_1 \text{ if } p = 0 \\ x_i - (x_i - x_{i-1})[i - 1 - p(n-1)] \quad \text{if } \dfrac{i-2}{n-1} < p \le \dfrac{i-1}{n-1} \quad \text{with } 2 \le i \le n \end{cases}$$

Note that a bias-correction is applied to estimate the confidence limits with the log-triangular distribution.

ACRONYMS

SSWD=Species Sensitivity Weighted Distribution

HC_{th}=Hazardous Concentration to the threshold of protection *th*

Sp=species or genus information

TW=Taxonomy or trophic level weight

R^2=Multiple R-square on the Quantile's data

KSpvalue= pvalue of the Kolmogorov-Smirnov goodness of fit test

GWM=Geometric Weighted Mean of the log-normal distribution

GWSD=Geometric Weighted Standard Deviation of the log-normal distribution

wm.lg=Weighted Mean of the log-normal distribution of the data (log10)

wsd.lg=Weighted Standard Deviation of the log-normal distribution of the data (log10)

GWMin=Geometric Min parameter of the Weighted log-triangular distribution

GWMax=Geometric Max parameter of the Weighted log-triangular distribution

GWMode=Geometric Mode parameter of the Weighted log-triangular distribution

wmin.lg=Min parameter of the weighted log-triangular distribution (log10)

wmax.lg=Max parameter of the weighted log-triangular distribution (log10)

wmode.lg=Mode parameter of the weighted log-triangular distribution (log10)

REFERENCES

1. Aldenberg T, and Slob W. 1993. Confidence limits for hazardous concentrations based on chronic NOEC toxicity data. *Ecotoxicol and Environ Saf* 25:48-63.
2. Forbes TL, Forbes VE. 1993. A critique of the use of distribution-based extrapolation models in ecotoxicology. *Funct Ecol* 7:249-254.
3. Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.
4. Versteeg DJ, Belanger SE, Carr GJ. 1999. Understanding single-species and model ecosystem sensitivity: data-based Comparison. *Environ Toxicol Chem* 18:1329-46.
5. National Institute of Public Health and the Environment in the Netherlands (RIVM). 1999. Environmental Risk Limits in the Netherlands. RIVM Report N°601640001.
6. van der Hoeven N. 2000. Estimating the 5-percentile of the species sensitivity distribution without any assumptions about the Distribution. *Ecotoxicology* 10:25-34.
7. Aldenberg T and Jaworska JS. 2000. Uncertainty of the hazardous concentration and proportion affected for normal species sensitivity distributions. *Ecotoxicol Environ Saf* 46:1:1-18.
8. Shao Quanxi. 2000. Estimation for hazardous concentrations based on NOEC toxicity data: an alternative approach. *Environmetrics* 11:583-595.
9. Wehrens R, Putter H, Buydens LMC. 2000. The bootstrap: a tutarial. *Chemometrics and intelligent laboratory systems* 54: 35-52.
10. Verdonck FAM, Jaworska J, Thas O, et al. 2001. Determining environmental standards using bootstrapping, bayesien and maximum likelihood techniques: a comparative study. *Analytica Chimica Acta* 446:429-438.
11. Posthuma L, II Suter GW II, Traas T (eds). 2002. *Species Sensitivity Distributions in Ecotoxicology*. Lewis Publishers, Boca Raton, FL, USA.
12. Forbes VE and Calow P. 2002. Species Sensitivity Distributions Revisited: A critical Appraisal. *Human and Ecological Risk Assessment* 8:3:473-492.
13. Wheeler JR, Grist EPM, Leung KMY and al. 2002. Species sensitivity distributions: data and model choice. *Marine Pollution Bulletin* 45:192-202.
14. Van Straalen NM. 2002. Threshold models for species sensitivity distributions applied to aquatic risk assessment for zinc. *Environ Toxicol Pharmacol* 11:167-172.
15. Grist EPM, Leung KMY, Wheelers JR and Crane M. 2002. Better bootstrap estimation of hazardous concentration threshold for aquatic assemblages. *Environ Toxicol Chem* 21:7:1151-1524.
16. European Commission. 2003. Technical Guidance Document in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. Part II. Edition 2. Office for Official Publications of the European Communities, Luxembourg.
17. Duboudin C, Ciffroy P, Magaud H. 2004. Effects of data manipulation and statistical methods on species sensitivity distribution. *Environ Toxicol Chem* 23:2:489-499.