UCL

Institut de statistique

Consulting report N° 01-01

Report on:

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

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 III

Choice of a weighing method in the aggregation of local monitoring data to a regional distribution

Bernadette Govaerts
Eric Lecoutre
Philippe Vanden Eeckaut

Institut de Statistique
Université Catholique de Louvain

1. Summary

The Institute of Statistics of UCL has developed during year 99, in the framework of a contract for Eurochlor, a methodology to summarise the concentration levels of chemical substances in the rivers of a region on the basis of monitoring data. The approach consists of aggregating the data observed in each sampling station to a regional statistical distribution from which any parameter of interest like a regional mean or percentile may be derived.

In the aggregation of local data to regional information, the methodology requires the definition of a weight for each sampling station. This weight is aimed at expressing the importance of each station in the studied region.

This paper proposes a methodology to derive such weight on the basis of hydrological and geographical information. It may be applied to any single river network from its sources to its estuary. It has the advantage to rely on very few information: the mean river flow in each sampling station and a map of the relative positions of the stations in the river network. This information allows to derive the weights through straightforward formulae.

1. Introduction

The Institute of Statistics of UCL has developed a methodology to summarize the concentration levels of chemical substances in the rivers of a region on the basis of monitoring data. This main idea of the approach is to build first a statistical distribution to summarise the values of concentrations observed in the sampling stations of the given region. This distribution may then be used to derive any parameter of interest like a regional mean, percentiles… The method has already been applied on several chlorine substance of the COMMPS data base (COMMPS (98)).

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.

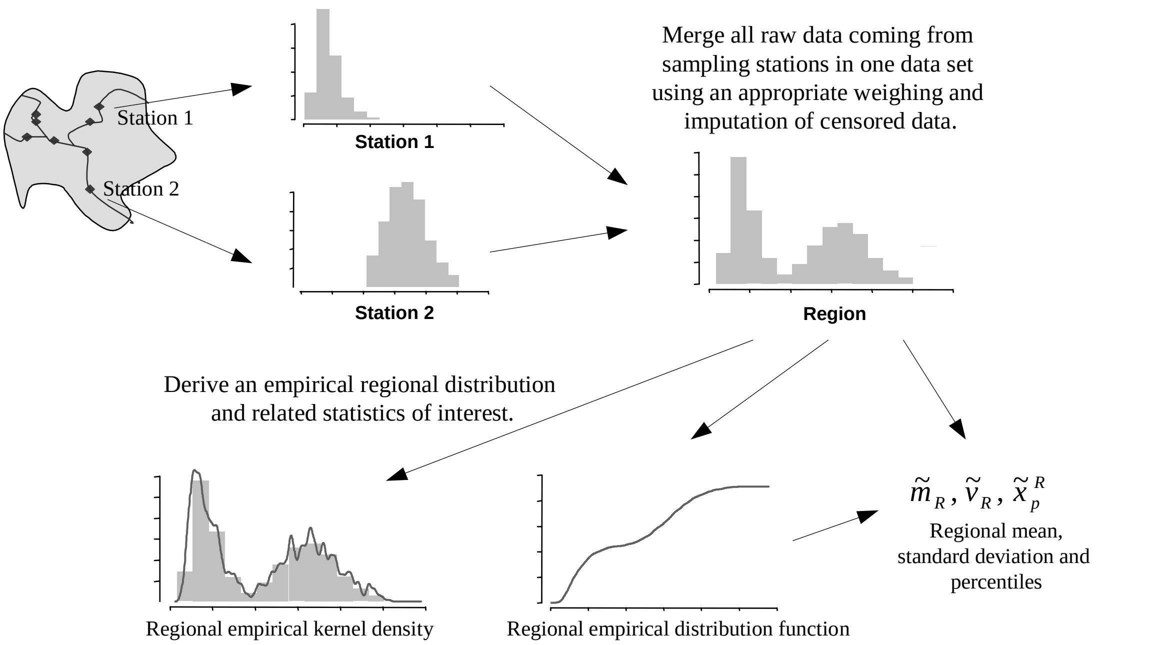


Figure 1 : Empirical method to derive a regional distribution

The two other approaches (one for complete data and another for summaries) are parametric since they assume 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.

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

In these non parametric and parametric approaches, all formulae proposed to derive empirical or parametric regional distributions from data collected in each sampling station include weights w_i. They reflect the importance of each sampling station in the region. Their choice is an open and difficult question since the notion of "importance" of a station depends on the use made of the regional distributions.

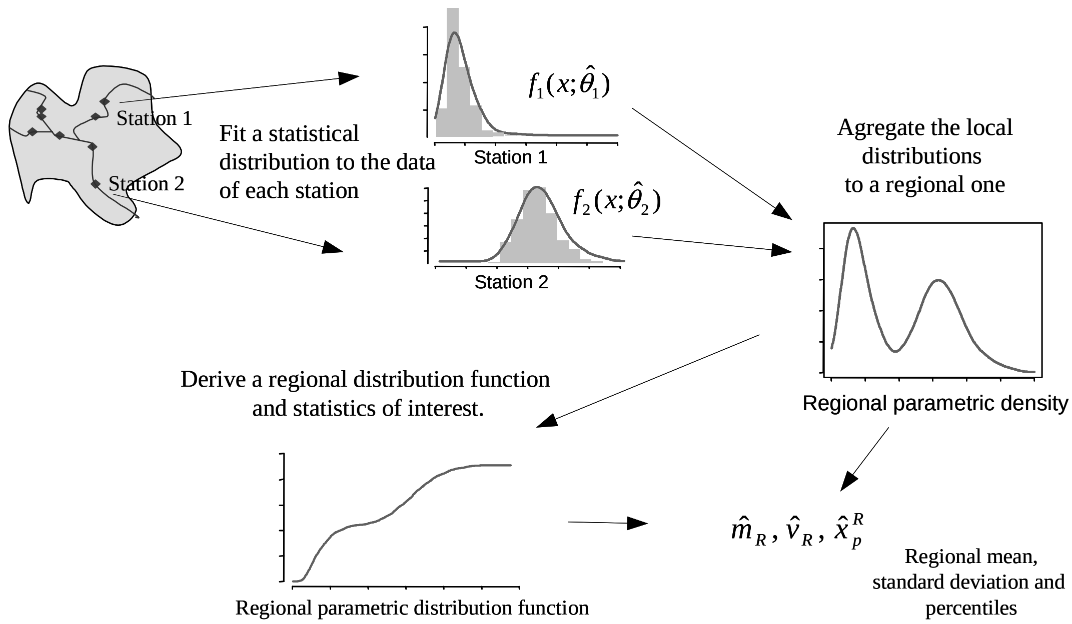


Figure 2 : Parametric method to derive a regional distribution

The simplest choice consists in giving an identical weight to each sampling station. This option may be adequate when the stations are assumed to be well spread over the region. The weights may also be fixed quite subjectively on the basis of economic, demographic or environmental criteria. For example, one might give to each sampling station a weight proportional to the size of the population living around it.

Another solution is to derive weights on the basis of hydrological and geographical information. A station located on a very small affluent of a river should not have the same weight than a station close to the estuary. Moreover, two stations close to each other should have their weights reduced accordingly.

This paper proposes a methodology to calculate weights for the different stations of a single river network from its various sources to the last station before the estuary. The methodology is the result of an iterative thinking on the subject made in two master thesis (Delande[2000] and Lecoutre[2001]) and in the paper Govaerts *et al* [2001]. Compared to these first attempts, the methodology proposed here has the advantage to be the simplest one and to rely on a very small number of hydrological information. In addition to the concentration observed in each sampling station, one should simply have a measure of the mean river flow in each station and a very elementary map of the river network giving for each station the station(s) which are directly above it (possibly located on different river affluents).

The reader interested by other developments available in the literature will find such information in Lecoutre [2001] or in Coffey [1987] but none of the solutions proposed there give a direct answer to the problem of weighing data before their aggregation over a region.

Section 2 of this paper presents the methodology in details starting with an elementary river with no affluent towards a general river network. Section 3 gives an illustration of the methodology on chloride (Cl-) concentration data collected on the Weser river in Germany. Finally section 4 discusses the practical issues of implementation of the methodology and its limitations.

2. A simple weighing approach

2.1 Weighing principle

The weighing principle proposed consists of deriving weights on the basis of hydrological and geographical information. As stated above, a station on a very small affluent of a river should not have the same weight than a station close to the estuary. Moreover, two stations close to each other should have their weights reduced accordingly.

This objective may be achieved by weighing each station according to the average portion of the river's water it is supposed to represent. The quantity of water in a river is not an information that is directly available. However, this unknown quantity may be related to other variables that are regularly measured such as the river flow. More precisely, upstream from a given station, the water quantity may be related to the size of the river catchment above the station. The latter may itself be related linearly to the mean flow of the river at the station of interest. As an example, Figure 3 shows the relation between the surface river catchment and average annual flows for 80 sampling stations situated on the Thames in England.

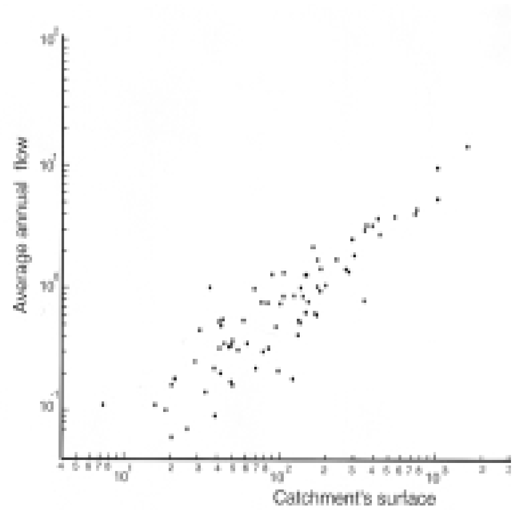


Figure 3 : Relation between average river flows and surface of station catchments on the Thames (source: Centre for Ecology and Hydrology - Wallingford - UK)

One may then interpret the mean flow at each sampling station as an indicator to express the importance of the river at that point. When two stations S_1 and S_2 are placed on a same river segment, the difference of flows (f_2-f_1) may be considered as a measure proportional to the quantity of water between the two stations.

2.2 *Derivation of weights for a river without affluent*

To see how these flows may be used to derive weights for the stations, let's take the easy case of a river without affluent with r sampling stations S_1, S_2 ... S_r and where the mean concentration and flows over a sampling period are respectively c_1, c_2 ... c_r and f_1, f_2 ... f_r. S_0 will denote the source of the river. At that point, $f_0=0$ and c_0 is defined as the background concentration at the river caption.

The average concentration over the river will be estimated assuming a linear evolution of the concentration between sampling stations. The positions of the stations on the river are defined by their respective mean flows (instead of their real geographic position) and express the importance of the river at each station. This is illustrated in Figure 4.

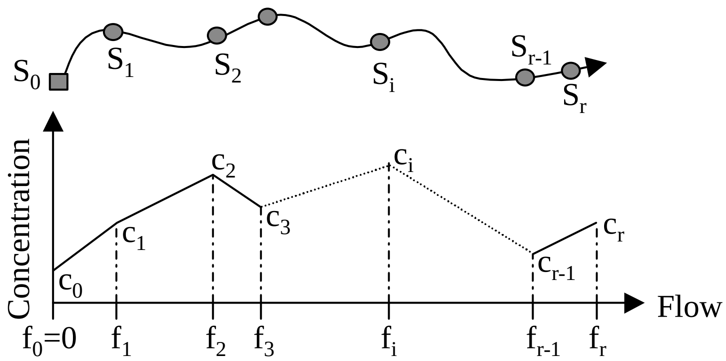


Figure 4 : Calculating an average concentration on a river by linear interpolation

Appendix 1 shows that, using linear interpolation, the mean concentration over the river may be estimated by the following formula:

$$\bar{c} = \frac{1}{f_r}\sum_{i=0}^{r} v_i c_i = \sum_{i=0}^{r} w_i c_i$$

where

$$v_0 = \frac{f_1}{2}, v_i = \frac{f_{i+1} - f_{i-1}}{2} \text{ for } i = 1,...r\text{-}1 \text{ and } v_r = \frac{f_r - f_{r-1}}{2}$$

Note that all the weights v_i must be divided by f_r to ensure that they sum to one.

The weights to be used in the aggregation of local information are then defined as:

$$w_i = \frac{v_i}{f_r}$$

This weighing may also be used in the aggregation of densities and distribution functions. Note also that this approach aggregates the information from the source of the river to the last station but says nothing about what happens downstream S_r. If S_r is not close to the estuary, the results will certainly not be representative of the whole river.

Below, we propose an additional hypothesis in this weighting. Since it is certainly not simple to choose a value for c_0 the background concentration at the river source, it is more realistic in practice to simply suppose that the concentration remains constant from the source to the first encountered sampling station (S_1 in this example). This is not a crucial hypothesis in the methodology but makes calculations simpler especially in the aggregation of densities and distribution functions.

With this further hypothesis, the weighing scheme becomes:

$$\bar{c} = \frac{1}{f_r}\sum_{i=1}^{r} v_i c_i = \sum_{i=1}^{r} w_i c_i$$

where

$$v_1 = \frac{f_1 + f_2}{2}, v_i = \frac{f_{i+1} - f_{i-1}}{2} \text{ for } i = 2,...r\text{-}1 \text{ and } v_r = \frac{f_r - f_{r-1}}{2}$$

2.3 Generalisation to rivers with one affluent

The generalisation of this methodology to a more complex river network is not straightforward and different solutions may be envisaged. This paper proposes probably the simplest possible solution that has the advantages of an easy implementation and to rely on limited hydrological and geographical information.

Let's first generalise the above interpolation approach to a river with one affluent as illustrated in Figure 5.

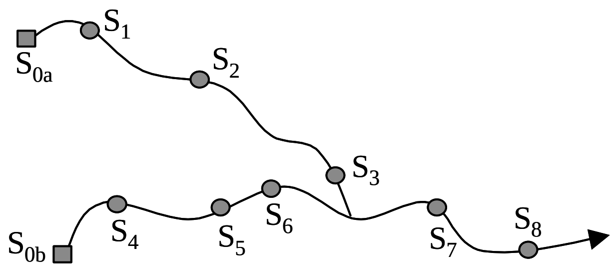


Figure 5 : Example of river with one affluent

The principle of the method is to redraw this simple network as a set of river segments with no affluent and then treat each segment with the method developed above. In this case, the river may be divided in three parts as shown in Figure 6.

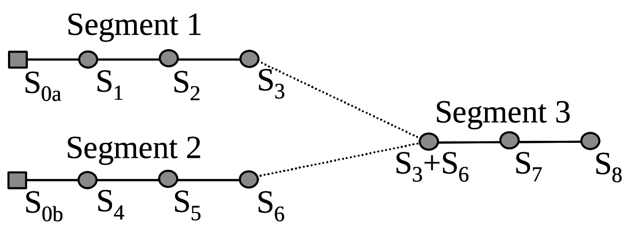


Figure 6 : Redrawing of a river with one affluent as 3 segments with no affluent.

For the river part between S_3, S_6 and S_7, one acts as if the junction was just after S_3 and S_6 and that S_3 and S_6 are linked to S_7 by a single river segment. On this third segment, S_3 and S_6 are taken as a single station where the mean flow is:

$$f_{36} = f_3 + f_6$$

and the mean concentration is:

$$c_{36} = \frac{f_3 c_3 + f_6 c_6}{f_3 + f_6}$$

The difference $f_7 - (f_3 + f_6)$ is then a way to quantify the importance of the river or of the portion of the river catchment between these three stations.

The weights are then derived by interpolation of the concentrations between sampling stations in each segment following the principle developed in the previous section. Appendix 2 shows that the following weight should be used for the sampling stations:

$$v_1 = \frac{f_2 + f_1}{2} \quad v_2 = \frac{f_3 - f_1}{2} \quad v_3 = \frac{f_3 f_7}{2(f_3 + f_6)} - \frac{f_2}{2}$$
$$v_4 = \frac{f_5 + f_4}{2} \quad v_5 = \frac{f_6 - f_4}{2} \quad v_6 = \frac{f_6 f_7}{2(f_3 + f_6)} - \frac{f_5}{2}$$
$$v_7 = \frac{f_8 - (f_3 + f_6)}{2} \qquad v_8 = \frac{f_8 - f_7}{2}$$

These formulae are rather intuitive. For most stations, the weight of a given station is close to the difference of flow between the downstream and upstream stations divided by two. For stations S_3 and S_6, the flow of the station below (S_7) is split in two parts:

$$\frac{f_3 f_7}{(f_3 + f_6)} \quad \text{and} \quad \frac{f_6 f_7}{(f_3 + f_6)}$$

to estimate the portion of f_7 to be attributed to each station. In S_7, the flow of the station upstream is (f_3+f_6) since, in the simplified network, these two stations have been aggregated.

2.4 General river network

For a general river network, the principles presented above may be applied. The network must first be redrawn as a set of river segments with no affluent and each segment treated independently by linear interpolation. Figure 7 gives an example of such a general network.

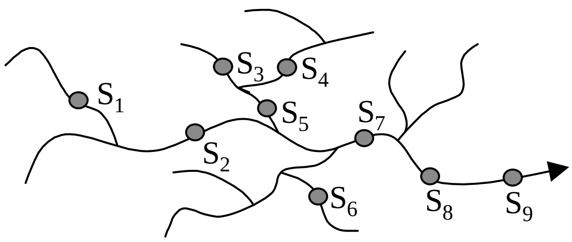


Figure 7 : Example of general river network

The simplified network will contain 6 segments and may be built iteratively from the bottom to the top of the river. In this example, it works as follow (see Figure 8):

1. S_9 is the bottom station and is put at the end of a first segment.
2. The river is then followed upstream and S_8 is encountered before any junction. S_8 is then set upstream on the same river segment than S_7.
3. Over S_8, a river junction is encountered but no station is present on the affluent. This affluent will be ignored since no information is available about it. It simply belongs to the portion of river catchment between S_7 and S_8.
4. Next, S_7 is set on the same segment than S_8 and S_9 following the principle of step 2.
5. Above S_7, there are three river segments with three sampling stations : S_2, S_5 and S_6 (the affluent with no station is ignored). The first simplified segment will then be closed by the aggregation of these three stations (S_{256} = S_2 + S_5 + S_6) and three new segments created with S_2, S_5 and S_6 as bottom stations.

6. This process is then reiterated from each of these 3 stations in order to build the three next simplified segments.
7. When there is no new station upstream a given station, the segment is ended by a source S_0.

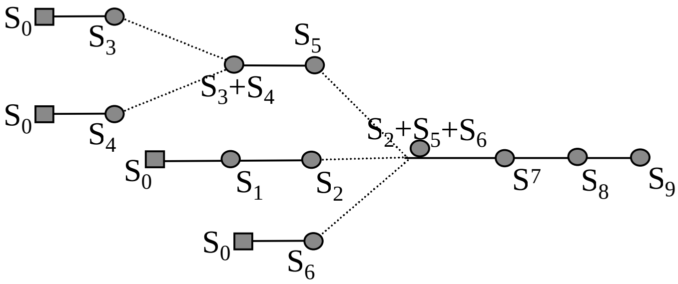


Figure 8 : Simplification of a general river network

The station weights for this network example can be derived easily following the logic discussed in section 2.3:

$$v_1 = \frac{f_2 + f_1}{2} \qquad v_2 = \frac{f_2 f_7}{2(f_2 + f_5 + f_6)} - \frac{f_1}{2}$$

$$v_3 = \frac{f_3 f_5}{2(f_3 + f_4)} + \frac{f_3}{2} \qquad v_4 = \frac{f_4 f_5}{2(f_3 + f_4)} + \frac{f_4}{2} \qquad v_5 = \frac{f_5 f_7}{2(f_2 + f_5 + f_6)} - \frac{f_3 + f_4}{2}$$

$$v_6 = \frac{f_6 f_7}{2(f_2 + f_5 + f_6)} + \frac{f_6}{2} \qquad v_7 = \frac{f_8 - (f_2 + f_5 + f_6)}{2} \qquad v_8 = \frac{f_9 - f_7}{2} \qquad v_9 = \frac{f_9 - f_8}{2}$$

Note that the sum of these weights is still f_9 the flow in the last station.

2.5 Automatic treatment of a general river network

The procedure presented in the previous section is a necessary step to explain the methodology but may still be implemented more simply. The idea is to rewrite the river network as a connection matrix and calculate the weights automatically from it.

For a river network with r sampling stations, the connection matrix is defined as a r $\times$ r square matrix made of zeros and ones. Each line and each column of the matrix is a sampling station and the element (i,j) of the matrix expresses the connection between a couple of sampling stations. More precisely, if station S_i is upstream station S_j and there is no station in between, x_{ij} is set to one. In all the other cases x_{ij}=0.

This gives the following matrix for the network given in figure 7:

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8 \\ S_9 \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 & S_8 & S_9 \end{array} \\
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array}$$

A general formula may then be used to calculate a weight w_i for each sampling station:

$$w_i = \frac{1}{f_r} v_i = \frac{1}{f_r} \left(\frac{1}{2} f_i \left(1 - \sum_{j=1}^{r} x_{ij} + I\{\sum_{j=1}^{r} x_{ji} = 0\} \right) + \frac{1}{2} \left(\sum_{j=1}^{r} f_j x_{ij} \right) \frac{f_i}{\sum_{j=1}^{r} x_{ij} \sum_{k=1}^{r} x_{kj} f_k} - \frac{1}{2} \sum_{j=1}^{r} f_j x_{ji} - \frac{1}{2} f_i \right)$$

$$\text{where} \quad I\{\sum_{j=1}^{r} x_{ji} = 0\} = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{r} x_{ji} = 0 \\ 0 & \text{otherwise} \end{cases}$$

3. Application of the methodology to Chloride data on the Weser (G)

This section gives an illustration of the weighing methodology presented in the previous section. It shows how to derive practically weighted regional statistics and distributions on chloride (Cl-) concentrations measured during year 1993 on the Weser in Germany.

3.1 The Weser data

The data have been extracted from a report prepared each year by the "Ministerium für Umwelt, Raumordnung und Landwirtschaft des Landes Nordrhein-Westfalen" on the evolution of the concentration of several chemical substances over the year in 17 sampling stations situated along the Weser and its affluents.

The configuration of the rive network is given in figure 9.

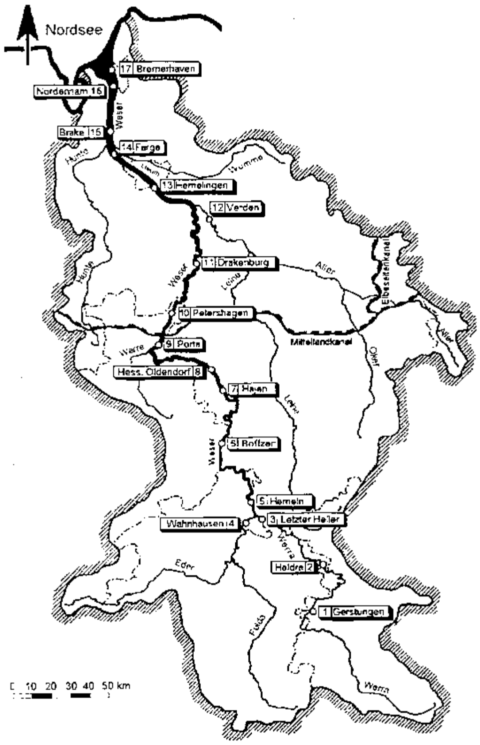


Figure 9 : Map of the Weser river network

Figure 8 gives a more schematic representation of the river :

Figure 10 : Schematic representation to the Weser river network.

The data collected each 2 weeks in the 17 sampling stations have the advantage to be completely listed in the report and to contain the river flows needed in our methodology. In our illustration, we have chosen to derive the regional statistics and distributions for the Chloride concentration because this series of data contain nearly no missing values and no data is censored. This makes the analysis simpler. The flows and concentrations data are listed in appendix 3. The analysis below focuses only on stations 1 to 13 because stations 14 to 17 are in the estuary of the Weser were the weighing methodology can no longer be applied directly. Indeed, in the estuary, the river flows are not any more representative of the importance of the river at each sampling station. They are either not provided in the report for these stations.

Delande [2000] has done a quite detailed study of the geographical and temporal evolution of these data and of the relation between Cl concentrations and river flows. Figure 11 gives a summary of the mean evolution of river flows and Cl concentrations over the 13 stations of the river network and over the sampling period. Table 1 gives also the mean flow, mean concentration and standard deviation of the concentration for each sampling station.

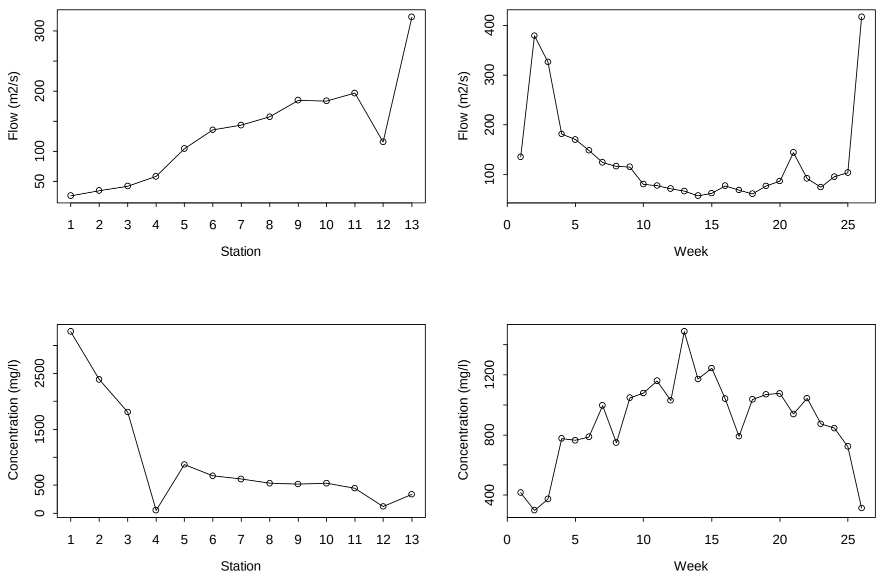


Figure 11 : Mean evolution of river flows and Cl concentration over the sampling stations and during the year in the Weser river in 1993.

These data show naturally that the river flow increase over the river (Station 12 is on an affluent) and that the flows are higher in the winter than in the summer. The concentrations have a tendency to decrease along the river and are lower in the two affluents (S_4 and S_{12}) than in the main river. In summer, concentrations increase as the flow decrease.

The reader interested by more details on these data should refer to Delande [2000].

3.2 Simplification of the river network and weights calculation

The derivation of weights for each sampling station requires first to redraw the river network as a set of segments without affluent. For the Weser, 5 segments are needed to redraw the network as shown in Figure 12.

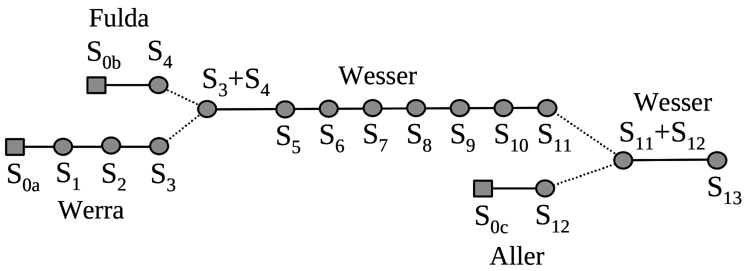


Figure 12 : Simplied presentation of the Weser river network

The weight formulae can then be derived easily following the same principles than above. We have :

$$w_i = \frac{v_i}{f_{13}}$$

$$v_1 = \frac{f_2 + f_1}{2} \quad v_i = \frac{f_{i+1} - f_{i-1}}{2} \quad \text{for i in 2,6,7,8,9,10}$$

$$v_3 = \frac{f_3 f_5}{2(f_3 + f_4)} - \frac{f_2}{2} \quad v_4 = \frac{f_4 f_5}{2(f_3 + f_4)} + \frac{f_4}{2} \quad v_5 = \frac{f_6}{2} - \frac{f_3 + f_4}{2}$$

$$v_{11} = \frac{f_{11} f_{13}}{2(f_{11} + f_{12})} - \frac{f_{10}}{2} \quad v_{12} = \frac{f_{12} f_{13}}{2(f_{11} + f_{12})} + \frac{f_{12}}{2} \quad v_{13} = \frac{f_{13}}{2} - \frac{f_{11} + f_{12}}{2}$$

Each flow f_i is defined as the mean of the flows observed in station S_i over the sampling period. These flows and related weights are given in Table 1.

			Observed Concentration	
Station	**Mean Flow**	**Weigth wi**	**Mean**	**Stand-Dev**
S1	26.1	0.0940	3245	949.2
S2	34.6	0.0247	2389	588.3
S3	42.1	0.0142	1808	551.9
S4	58.4	0.1842	52	11.4
S5	104.5	0.0544	868	286.0
S6	135.7	0.0605	665	218.5
S7	143.6	0.0335	611	200.4
S8	157.3	0.0639	536	176.0
S9	184.9	0.0406	518	197.3
S10	183.6	0.0184	534	163.7
S11	196.8	0.0308	443	152.9
S12	115.7	0.3643	119	19.5
S13	323.2	0.0165	334	140.1

Table 1 : Stations weights and summary statistics for the Weser river

3.3 Estimation of local distributions for the concentration

The derivation of the regional information will be done here by the parametric method which requires first the fitting of a local parametric distribution to the data of each sampling station. For these data, a normal distribution has been chosen as local parametric distribution on the basis of a statistical analysis using qqplots and Kolmogorov-Smirnov tests. Appendix 4 compares for each sampling station the qq plot of the data when they are supposed to be normal and lognormal. If the data follow the distribution under test in the plot, the data points should lie closely to a strait line. The plots show quite clearly that most normal qq plots look more linear than lognormal ones. On each qqplot, the p-value of the Kolmogorov-Smirnov test for normality or lognormality are also given and show that p-values are nearly always higher for normal qqplot that for lognormal ones. This means that for most of the stations (out of station 13) data are closer to normality than to lognormality.

The estimation of local normal distribution is straightforward. The two parameters of the normal μ and σ^2 are simply estimated by the arithmetic mean and sample variance of the station data. These are given in Table 1. Note that, even if locally the distributions are close to normality, over the region, the concentration varies on a very wide range and should be treated on a log scale.

3.4 Comparison of weighted and non weighted regional statistics and distributions

This section shows how to derive weighted and non weighted regional information from local distributions fitted to station data. Figure 12 shows first the regional density and distribution functions. They are derived using the following formulae (see also Govaerts et all [2001])

$$\hat{f}_R(x) = \sum_{i=1}^{r} w_i\, f_i(x;\hat{\theta}_i) = \sum_{i=1}^{r} w_i \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{(x-\hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\right)$$

$$\hat{F}_R(x) = \sum_{i=1}^{r} w_i F_i(x;\hat{\theta}_i) = \sum_{i=1}^{r} w_i \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{(x-\hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\right) dx$$

where w_i is either 1/13 if the station are not weighted or the weights given in Table 1. A big difference between the two approaches comes also from the treatments of the sources. In the non weighted approach, the 13 stations are the only information used in the aggregations. However, in the weighted approach, sources are in a way included as stations in the analysis through the weights of the stations just below them.

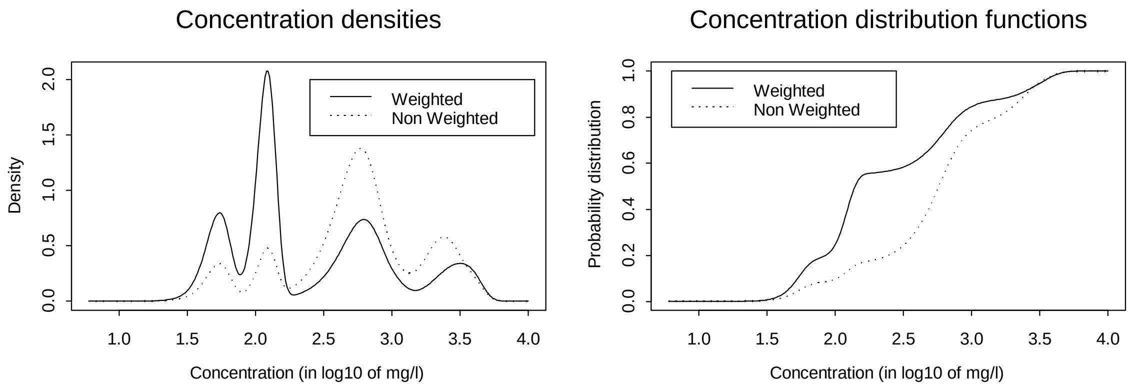


Figure 13 : Comparison of weighted and non weighted densities and distribution functions

The difference between the two results can be explained first by the introduction of the sources in the systems which have the tendency to decrease concentration by giving more weights to small concentrations. It comes also from the non equal weights attributed to the stations. As an example, station 12 has a weight which is 10 times the weight of station 11 due to the fact that the river branch over which S_{12} has a very high flow compare to the rest of the river.

The regional mean, variance and percentiles can be derived from the following formulae :

$$\hat{m}_R = \sum_{i=1}^{r} w_i \hat{\mu}_i \qquad \hat{v}_R^2 = \sum_{i=1}^{r} w_i (\hat{\sigma}_i^2 + (\hat{\mu}_i - \hat{m}_R)^2) \quad \text{and} \quad x_p^R = \min\{x \in R^+ \,|\, \hat{F}_R(x) \geq p\}$$

Table 2 compares the values obtained in the weighted and non weighted cases.

	Weighted	**Non Weighted**
Mean	635	932
Standard deviation	992	1007
P50	139	573
P75	655	1031
P90	2198	2494
P95	3210	3140

Table 2 : Comparison of Weser regional weighted and non weighted statistics

Note that all these calculations have also been done using lognormal local distributions and the results are very similar. This is expected because at a local level, concentrations are very similar compared to the variation observed over the region.

4. Conclusion

This paper presents a simple and straightforward methodology to weight sampling stations data in their aggregation to regional information. This methodology has been developed in the framework of a more general research project realised by the Institute of Statistics of UCL for Eurochlor but can certainly find applications in many contexts where river data must be aggregated. This is in particular true in water risk assessment where it is usually supposed that the stations are well distributed on the river network which bypass the problem of weighing (see COMMPS (1998) or WFD-GM.(2001)).

The direct application of the methodology to current international databases is of course not straightforward. Such databases do usually not supply the configuration of the river networks where sampling stations are present and no information is available on river flows. This last information is even not necessary available in the hydrological national databases since hydrological data are usually not collected by the same organisations than monitoring data. As a result, when river flows are available, they have often not been measured at the same stations than the concentration levels. Moreover, at a national or international level, the data collected are coming from rivers but also from lakes, canals, estuaries, seas... and the methodology proposed here is limited to river networks. Other weighing schemes should be developed for these other classes of surface water and also for the relative weighing of one to the other.

A validation of the methodology needs also to be done from a hydrologic viewpoint in order to assess if the approximation of the size of the river catchment before a sampling station with a mean river flow is acceptable for all types of rivers. The validation of this hypothesis does not call into question the methodology since we can also apply it directly on the size of the river catchments when this information is available.

Appendix 1: Derivation of weights for a river without affluents

The mean concentration over the river may be defined as the weighted sum of mean concentrations on each river elementary segment $S_i \rightarrow S_{i+1}$. This mean is given by :

$$\frac{c_i + c_{i+1}}{2}$$

Each river segment is weighted by the difference of flows between the two stations which is a measure of the importance of the river between them or more precisely the part of the river catchment between the two points. The weights are normalized to sum to one by being divided by $1/f_r$. This gives :

$$\bar{c} = \frac{1}{f_r}\left(\frac{c_0 + c_1}{2} f_1 + \frac{c_1 + c_2}{2}(f_2 - f_1) + \frac{c_2 + c_3}{2}(f_3 - f_2) + \ldots + \frac{c_{r-1} + c_r}{2}(f_r - f_{r-1})\right)$$

This expression may be reorganized to get a weighted sum of local concentrations:

$$\bar{c} = \frac{1}{f_r}\left(\frac{f_1}{2} c_0 + \sum_{i=1}^{r-1} \frac{f_{i+1} - f_{i-1}}{2} c_i + \frac{f_r - f_{r-1}}{2} c_r\right)$$

Appendix 2: Derivation of weights for a river with one affluent

The mean concentration over the river is derived by interpolation on each simplified segment as if it was a river without affluent:

$$\bar{c} = \frac{1}{f_r}\left(\frac{c_{0a} + c_1}{2} f_1 + \frac{c_1 + c_2}{2}(f_2 - f_1) + \frac{c_2 + c_3}{2}(f_3 - f_2)\right. \qquad \text{Segment 1}$$

$$+ \frac{c_{0b} + c_4}{2} f_4 + \frac{c_4 + c_5}{2}(f_5 - f_4) + \frac{c_5 + c_6}{2}(f_6 - f_5) \qquad \text{Segment 2}$$

$$\left. + \frac{c_{36} + c_7}{2}(f_7 - f_{36}) + \frac{c_7 + c_8}{2}(f_8 - f_7) + \frac{c_8 + c_9}{2}(f_9 - f_8)\right) \qquad \text{Segment 3}$$

were:

$$f_{36} = f_3 + f_6 \quad \text{and} \quad c_{36} = \frac{f_3 c_3 + f_6 c_6}{f_3 + f_6}$$

As in appendix 1, this expression may be reorganised to get weights for each station. We have:

$$v_{0a} = \frac{f_1}{2} \quad v_1 = \frac{f_2}{2} \quad v_2 = \frac{f_3 - f_1}{2}$$

$$v_3 = \frac{f_3 - f_2}{2} + \frac{f_7 - (f_3 + f_6)}{2} \times \frac{f_3}{f_3 + f_6} = \frac{f_3 f_7}{2(f_3 + f_6)} - \frac{f_2}{2}$$

v_{0a}, v_4, v_5, v_6 have similar expressions :

$$v_{0b} = \frac{f_4}{2} \quad v_4 = \frac{f_5}{2} \quad v_5 = \frac{f_6 - f_4}{2} \quad v_6 = \frac{f_6 f_7}{2(f_3 + f_6)} - \frac{f_5}{2}$$

Finally, w_6,and w_8 are obtained as follow :

$$v_7 = \frac{f_7 - (f_3 + f_6)}{2} + \frac{f_8 - f_7}{2} = \frac{f_8 - (f_3 + f_6)}{2} \qquad v_8 = \frac{f_8 - f_7}{2}$$

If we now suppose that the concentration remains constant from the sources to the sampling station just below ($c_{0a} = c_1$ and $c_{0b} = c_4$) than v_{0a} and v_{0b} disappear and v_1 and v_4 become :

$$v_1 = \frac{f_1 + f_2}{2} \quad v_4 = \frac{f_5 + f_4}{2}$$

Appendix 3 : Weser data

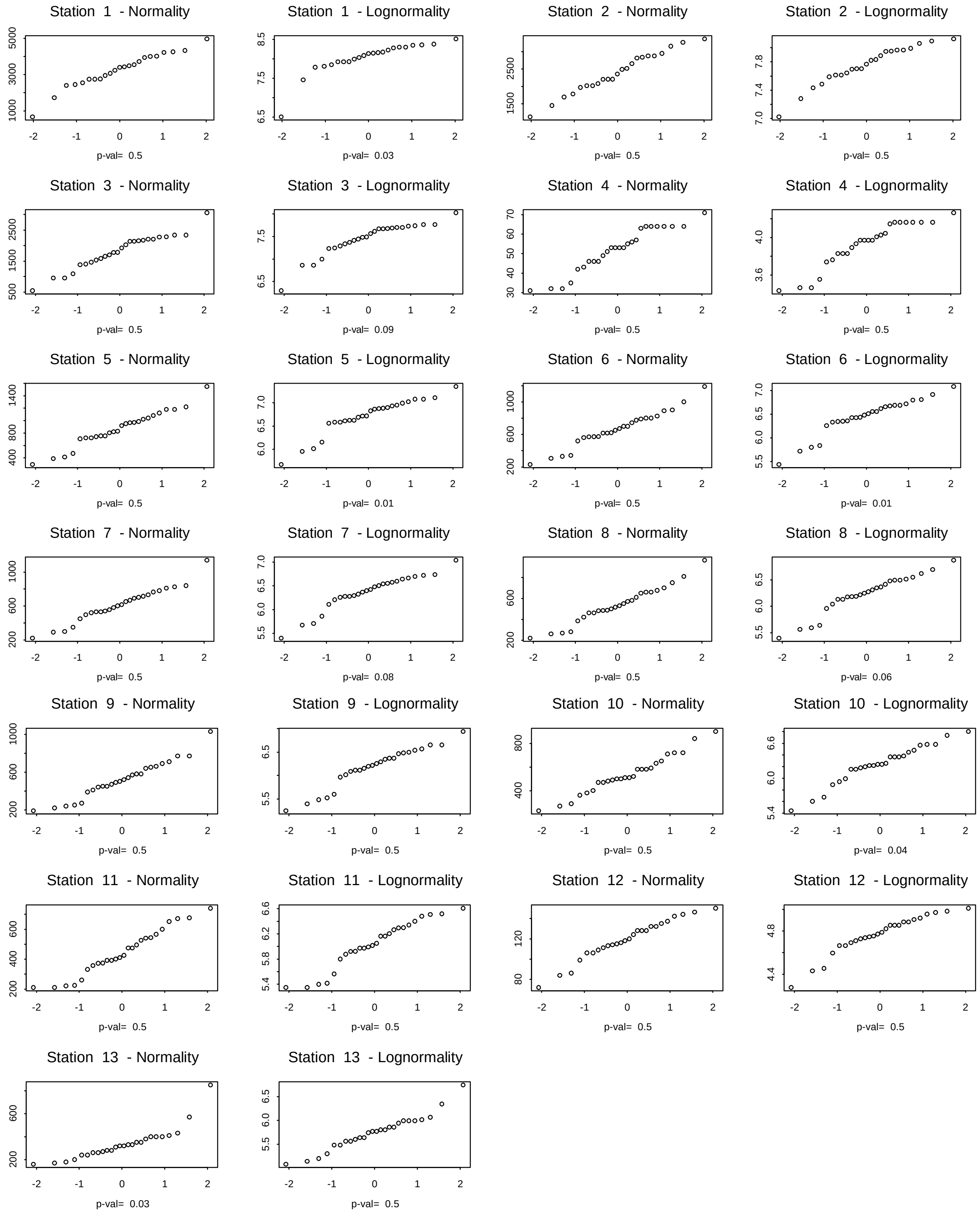
River Flows (m^2/sec)

Period	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
1	29.4	36.5	50.6	60.7	102.8	136.2	146.2	157.1	187.0	186.2	202.7	125.3	339.8
2	82.6	92.8	117.3	236.2	379.5	429.0	448.1	492.3	553.6	537.6	545.2	223.9	794.8
3	53.6	92.8	80.6	153.1	248.3	325.6	354.7	398.5	491.8	478.1	523.3	237.1	807.1
4	29.2	41.1	54.8	66.9	129.9	180.9	193.5	216.5	259.1	250.0	283.1	173.1	482.1
5	24.4	37.5	52.9	54.2	117.4	165.0	179.3	196.7	238.8	232.7	266.4	172.8	473.8
6	38.1	47.1	60.6	54.4	123.1	159.7	167.7	177.6	201.4	196.9	215.6	128.4	362.5
7	31.7	41.4	52.9	39.4	101.4	135.3	141.4	148.0	169.3	166.0	180.2	108.9	302.6
8	23.5	32.4	42.5	42.7	91.9	123.8	132.2	139.5	161.6	161.5	176.4	101.5	287.7
9	21.8	29.3	38.3	44.9	87.3	122.5	130.4	138.5	164.4	164.0	176.5	96.7	284.4
10	15.8	23.3	28.0	31.8	60.7	86.6	93.7	100.0	116.1	109.4	117.4	70.8	193.9
11	14.7	21.9	27.3	27.9	56.3	79.6	85.8	93.5	113.0	108.0	115.0	74.7	194.2
12	13.9	20.0	27.2	25.1	54.8	76.5	81.4	89.1	101.3	94.4	99.1	70.7	175.6
13	13.3	17.7	23.4	26.8	51.3	70.0	75.2	82.1	93.7	92.2	91.0	68.0	162.8
14	10.4	14.9	18.4	28.5	48.5	63.7	68.7	73.8	81.5	78.9	75.7	51.7	131.2
15	17.7	16.4	20.3	26.4	47.1	65.7	70.4	76.7	87.3	86.8	87.3	60.6	150.2
16	15.0	18.7	23.1	28.1	51.8	71.8	76.6	80.8	103.6	107.5	112.3	99.8	216.1
17	15.1	19.0	22.8	24.6	48.2	65.2	70.3	75.5	90.2	91.8	94.9	87.9	187.9
18	10.8	14.4	16.8	30.2	48.4	64.9	69.5	73.4	81.6	81.6	83.3	68.1	154.6
19	12.4	16.5	19.2	29.4	50.0	69.7	75.9	87.9	108.4	112.9	121.9	87.1	208.8
20	11.5	15.1	18.3	37.6	56.6	76.0	82.0	91.9	123.0	126.0	145.1	102.9	245.3
21	12.8	18.5	25.5	52.0	81.2	128.3	138.5	161.7	215.4	223.1	266.1	135.4	417.3
22	11.9	16.7	20.7	35.9	59.2	89.3	96.1	104.1	126.4	129.8	150.2	98.6	258.8
23	11.9	16.7	20.3	29.7	52.6	74.0	78.4	83.9	100.7	101.0	116.5	80.6	202.4
24	15.2	20.8	27.9	35.0	66.9	93.3	99.3	106.9	129.1	132.1	148.6	106.0	260.2
25	23.0	27.3	32.4	46.7	60.5	98.4	100.3	111.5	141.6	144.1	164.4	114.0	284.5
26	120.0	151.4	173.1	249.6	441.2	476.9	477.9	532.9	567.5	580.1	557.9	264.8	823.9

Cl Concentration (mg/l)

Period	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
1		1450	952	53	470	330	350	268	250	360	225	106	180
2			952	53	384	342	290	280	240	270	221	86	170
3		1120	1090	31	410	304	300	260	220	290	210	84	160
4	2750	2080	1530	32	704	616	496	420	390	400	330	106	240
5	2550	1970	1400	51	740	560	520	460	440	470	392	99	280
6	2450	1780	1460	46	825	615	650	550	470	580	410	111	280
7	2750	2020	1780	43	1080	890	810	750	770	840	670	120	430
8	2400	1690	1380	53	720	568	530	486	520	500	424	114	350
9	3400	2820	2130	32	984	824	690	530	710	580	474	109	320
10	3240	2520	2150	57	1220	1000	714	660	640	720	540	146	410
11	3720	3160	2270	64	1180	775	825	675	660	630	600	128	400
12	3540	2200	1920	64	1040	790	730	660	690	710	525	132	380
13	4970	3270	3060	63	1550	1190	1140	965	1030	900	675	137	400
14	4260	2660	2200	64	1020	800	780	700	650	650	740	150	570
15	4000	3370	2340	71	1180	900	840	810	770	720	650	135	400
16	3490	2360	2200	64	1120	800	765	650	570	490	565	116	350
17	2760		1650	64	950	700	666	580	580	590	494	128	320
18	3430	2490	2340	55	970	700	700	610	580	580	544	142	330
19	4330	2840	2170	64	916	670	580	516	490	500	400	124	310
20	4220	2880	2130	53	964	744	614	570	540	520	372	115	260
21	3940	2880	2020	42	750	520	450	386	270	380	260	113	200
22	4010	2950	2280	46	820	650	600	500	450	510	373	132	260
23	2950	2210	1770	46	800	620	556	480	500	480	474	144	330
24	3070	2210	1580	49	750	570	530	484	450	510	392	118	270
25	1730	2020	1700	56	720	574	540	460	410	470	356	128	240
26	666		541	35	292	230	220	220	190	230	210	72	850

Appendix 4 : Goodness of fit tests for Weser data



Bibliography

Coffey W. (1981), *Geography : towards a general spatial approach*, chapter 8, Methuen, New York.

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.

Delande L. (2000), *Utilisation d'informations géographiques pour le calcul de distributions régionales de concentration de produits chimiques dans les rivières européennes*, Mémoire de licence en sciences mathématique, Université catholique de Louvain.

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.

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.

MURL (1993), *Arbeitsgemeinschaft zur Reinhaltung Der Weser*, Ministerium fur Umwelt, Raumornung und Lanwirtschaft des Landes Nordrheim-Westfalen, Guterbericht.

WFD-GW (2001), *The EU Water Framework Directive : Statistical aspects of the identification of ground population trends and aggregation of minitoring results.* European commission report.