These

Subthese

Author: Mr These

September 1, 2022

Keywords:
data, energy, blockchain,

Conducted under the supervision of:

prof 1	-	affiliation prof 1
prof 2	-	affiliation prof 2
prof 3	-	affiliation prof 3
prof 4	-	affiliation prof 4
prof 5	-	affiliation prof 5

Contents

1 Chapter 1: Introduction **9**

2 Chapter 2 : Change-point detection state of the art **10**
2.1 Models . . . 11
2.1.1 Parametric setting . . . 12
2.1.2 Non-Parametric setting . . . 12
2.2 Searching for change points . . . 13
2.2.1 Optimal segmentation approach . . . 13
2.2.2 PELT algorithm . . . 14
2.3 Estimation of the number of changes . . . 14
2.3.1 Different heuristics . . . 14
2.3.2 Exploratory approach: CROPS algorithm . . . 14
2.4 Extension to the multivariate case . . . 16
2.4.1 Fully multivariate detection . . . 16
2.4.2 Subset multivariate detection . . . 16

3 Chapter 3 : Problematic and data **17**
3.1 General overview . . . 19
3.1.1 Emission source . . . 19
3.1.2 Substance diffusion . . . 20
3.1.3 Impregnation levels and potential adverse effects . . . 23
3.2 Environmental data specificity . . . 24
3.2.1 Inherent data characteristics . . . 25
3.2.2 Heterogeneity in distribution and sampling rates . . . 26
3.2.3 Representation . . . 27

4 Chapter 4 : Change point detection in left censored observations **33**
4.1 Model . . . 33
4.1.1 Estimating change-points locations with $K^\star$ known . . . 34
4.1.2 Estimating change-points locations with $K^\star$ unknown . . . 36
4.2 Estimation procedure . . . 37
4.2.1 Estimating the λ's . . . 37
4.2.2 Estimation procedure for $K^\star$, $\boldsymbol{t}^\star$ and $\boldsymbol{\lambda}^\star$ . . . 37
4.3 Simulation study . . . 38
4.3.1 Tuning hyper-parameters of the Newton-Raphson method 39
4.3.2 Tuning the parameters of the change-point detection method . . . 41

5 Chapter 5 : Case study **46**
5.1 Data collection procedure and associated generative model . . . 48
5.1.1 Monitoring stations network . . . 48

5.1.2 Data collection 49
5.1.3 A piece-wise stationary model for the coarse-grain time series 50
5.2 Methods 50
5.2.1 Piece-wise stationary model estimation via change-point detection 50
5.2.2 Spatial clustering 52
5.2.3 Anomaly detection 53
5.3 Data presentation 55
5.3.1 Time period and geographical area selection 55
5.3.2 Graphical representation of the station network 56
5.4 Results 57
5.4.1 Temporal segmentation 57
5.4.2 Spatial segmentation 59
5.4.3 Anomalous cluster identification 61

6 Conclusion **64**
6.1 Summary 64
6.2 Openings 64

Appendices **73**

A Proofs of Chapter 4 **74**
A.1 On the convergence of $\hat{\lambda}$ 74
A.2 Convergence of $\hat{K}$ and $\hat{t}$ 76
A.3 Verifying PELT assumptions 77

B Complement of Chapter 5 **78**
B.1 Simulation on the convergence of σ 78
B.2 Clustering algorithms 79
B.3 Modified empirical Wasserstein distance 82
B.4 Supplementary Figures 82
B.4.1 Regional map of crops 82
B.4.2 Prosulfocarb sales 83
B.4.3 All elbow methods figures 83

List of Tables

4.1 Choice of initialisitation value: simulation results for $n = 20$. . . 41
4.2 Choice of initialisitation value: simulation results for $n = 100$. . 42
4.3 Choice of maximum number of iteration N_{max}: simulation results. 42
4.4 Number of correct estimations of K over $N = 100$ samples for both methods for different $\alpha\%$ censorship rates. 44

List of Algorithms

1 Optimal partition algorithm: 14
2 PELT algorithm 15
3 CROPS algorithm 15
4 PELT algorithm 39
5 Clustering with greedy method: 80
6 Clustering by dynamic programming: 81

List of Figures

3.1 Three main compartments of a complete monitoring procedure. 19
3.2 Stations monitoring water suface in the Centre-Val de Loire french region. Two different geographical scales are represented. The underlying hydrographic network linking all stations is plotted on the left, the stations are colored according to their hydroecoregion on the right. 22
3.3 All active stations measuring air quality on March the 1st of 2021 coupled with meteorological stations active that day. The main wind direction and speed measured that day is mapped with the red arrows. 23
3.4 Data sets selected in this work. 25
3.5 Censorship illustration. The top Figure sums up the limits of measurements effects. The bottom Figure shows the consequences of censorship on the samples of a station located in the Centre-Val de Loire region. This station changed its equipment in 2016, the LOQ change values. 26
3.6 Histogram of all measurements of prosulfocarbe made between 2017-10-01 and 2018-02-01 in the Centre-Val de Loire region. . . 27
3.7 Spatial and temporal heterogeneity in sampling. The Figures on the left represent all the samples of two neighbouring stations. The map on the right shows the position of those stations. . . . 28
3.8 Spatial and temporal heterogeneity in distribution. The Figures on the left represent all the samples of two stations. The map on the right shows the position of those stations. 29
3.9 Time series plot of HER 8,5 and 4 daily maximum concentrations. The log scale was used for a easier visualization. 30
3.10 Spatial maps in time. The prosulfocarbe's quantification rate of each station was computed for each season of 2017 and 2018. 31
3.11 Space (1-D)/Time plots. Grey tiles correspond to location and moment were no sample were collected. 32

4.1 Plot of the cost function values against λ values when all observations are censored. It is represented for several σ values. The sample consists in 100 values of threshold $a = 0.1$. 38
4.2 Choice of the minimal segment length: simulation results. Our method performance is Illustrated with the red line, the *Multrank* method is drawn in blue. The results are illustrated for several censorship thresholds and the different minimal segment lengths used were $5, 10, 25, 50$ and 75 observations. 43
4.3 Example of simulated signal with $(\lambda_1 = 1, \lambda_2 = 4, \lambda_3 = 0.5, \lambda_4 = 5, \lambda_5 = 1)$, $\sigma = 0.5$, $n = 400$, $K = 4$, $(p_1 = 80, p_2 = 160, p_3 = 240, p_4 = 320)$ and $\alpha = 50\%$. 44

4.4 Precision of the estimated change-points for both methods. . . . 45

5.1 Distribution of the number of measurements per station. 56
5.2 Plot of daily maximum concentrations 57
5.3 Map of the non connex components in the station graph. 58
5.4 Plot of successive $\widehat{\sigma}$ values. We stopped the to iterate when the $|\widehat{\sigma}_b - \widehat{\sigma}_b| \leq 10^{-3}$ 59
5.5 Best segmentation found by the change-point detection procedure with CROPS-based penalty tuning. The dates of the breaks are : October 20, 2012; May 25, 2016; October 13, 2016; February 7, 2017; October 5, 2017; January 19, 2018; October 5, 2018; January 18, 2019; October 11, 2019; May 6, 2020; October 7, 2020; December 20, 2020; July 27, 2021.The black rectangle corresponds to the selected temporal segment in section 5.4.2 . . 60
5.6 Map of geographical clusters. 61
5.7 Clusters pareto front. 62
5.8 Mapped pareto front. 63

A.1 Iteration function plot. The values of y_i used was drawned from a 1000000 size samples of Weibull realisations with parameters $(\lambda^\star, \sigma)$ 75
A.2 $\partial^2\mathcal{L}(\lambda; y, \sigma)$ plot. The values of y_i used was drawned from a 1000000 size samples of Weibull realisations with parameters $(\lambda^\star, \sigma)$. 76

B.1 Scenarios with $\sigma = 0.4$. 78
B.2 Scenarios with $\sigma = 0.8$. 79
B.3 Example of three stations data. The data were simulated. 84
B.4 Example of modified c.d.f. for the Wasserstein distance. 85
B.5 Wheat (in yellow) and barley (in red) crops location in Centre-Val de Loire 85
B.6 Prosulfocarb sales between 2008 and 2017 in the Centre-Val de Loire region 86
B.7 Elbow method selecting the optimal segmentation of the full signal $\overline{\mathcal{D}}$. 86
B.8 Elbow method for the spatial clustering. 87

1. Chapter 1: Introduction

2. Chapter 2 : Change-point detection state of the art

Contents

2.1 Models 11

2.1.1 Parametric setting 12

2.1.2 Non-Parametric setting 12

2.2 Searching for change points 13

2.2.1 Optimal segmentation approach 13

2.2.2 PELT algorithm 14

2.3 Estimation of the number of changes 14

2.3.1 Different heuristics 14

Methods for penalty calibration 14

Methods using the optimal partitioning algorithm 14

2.3.2 Exploratory approach: CROPS algorithm 14

2.4 Extension to the multivariate case 16

2.4.1 Fully multivariate detection 16

2.4.2 Subset multivariate detection 16

As discussed in Chapter 1, it is critical to determine time periods during which the properties of the concentration signal are constant. This is a mathematical problem that has been extensively addressed in the literature in the form of breakpoint detection. There are two types of breakpoint detection algorithms:

- **Off-line methods:** There are numerous reviews of this topic in the literature, among the most extensive are [9, 92]. In these methods, one works with the entire data signal. Breakpoints are identified by following the signal from the most recent date to the earliest date. There are numerous applications of these methods in different application areas [16, 54, 87, 58]

- **On-line methods:** a detailed description of these methods can be found in [9]. The difference with the methods of the previous point is that the detection of a breakpoint is done when the data is read. The goal is to detect as quickly as possible a point in time during the reading where the characteristics of the signal have changed.

This work will focus only on the off-line methods that we will develop in the following sections. This choice was motivated by the operational speed of data samplng and storing on which this work is based. Clearly, on-line detection methods would be of interest to the Anses if the goal were to act urgently and investigate any potentially anomalous signal as soon as it is detected. However, the agency's missions are longer term in nature. At the time this work began, data were being collected annually, so the use of online methods was inappropriate. Nevertheless, for readers who wish to refer to it, we can state that there is no shortage of online detection methods in the literature [62, 58, 43, 79, 56].
Breakpoint detection methods were excluded from this work and therefore are not be presented in this state of the art. Although these methods are of interest to the task set in this manuscript. The focus of this work was to explore frequentist models and their adaptation to the specificities of concentration data (see Chapter 3). However, numerous examples of application of bayesian methods can be found in the literature [82, 91, 1, 102, 59].
The outline of the state of the art is strongly based on the nomenclature established in (Truong). Section 1 describes the modelling used in the theory of breakpoint detection theory in parametric and nonparametric frameworks will be described in Section 1. The breakpoint search methods are then discussed in Section 2. The various existing methods for calibrating the penalty for the case where the number of breaks is unknown are presented in Section 3. Although the bulk of our work deals with univariate data, we will extend the methods discussed to the case where the variables are multivariate.

2.1. Models

Several terms are introduced and will be retained in this section. We consider a signal consisting of observations $\boldsymbol{y} = (y_1, ..., y_n)$ which are the realisations of random variables $Y_1, ..., Y_n$. This signal is assumed to be piecewise stationary. Its properties are constant over parts, which we will call segments, and change at times $t_1^* < ... < t_k^* < ... < t_{K^*}^*$. A segment of the signal from the u-th coordinate to the v-th is noted $y_{u:v}$. Following the convention, let $t_0 = 0$ and $t_{K^*+1} = n$. In our context, the purpose of breakpoint detection is to estimate the positions t_k^* and the number of breaks K^* when it is unknown.
The search for a segmentation $\mathcal{T}_{K^*}$ of a signal $\boldsymbol{y}$ into K^*+1 segments can be formally described as an optimization problem. More precisely, one tries to minimise a quantitative criterion. Without loss of generality, the criterion chosen in our case corresponds to a cost $\mathcal{C}(\mathcal{T}, \boldsymbol{y})$ and, more precisely, to the sum of the cost of each segment defined by $\mathcal{T}_{K^*}$:

$$\mathcal{C}(\mathcal{T}_{K^*}, \boldsymbol{y}) = \sum_{k=0}^{K^*} W(y_{t_k^*+1:t_{k+1}^*}), \tag{2.1}$$

where $W(y_{t_k^*+1:t_{k+1}^*})$ denotes the cost of the kth segment. The ultimate goal is to find the segmentation that minimizes 2.1. The function chosen to model the cost of a segment W determines the types of changes detected. Its choice can also depend on whether a parametric of non parametric settings is desirable. It is assumed that K^* is always assumed to be unknown. It will sometimes be necessary to assume that K^* is known in order to introduce certain algorithms. In this case, it is explicitly stated that this will be the case. the knowledge of the

number of breakpoints change the optimization problem. In the case, where K^* is known the problem consists in solving:

$$\min_{|\mathcal{T}|=K^*} \mathcal{C}(\mathcal{T}, \boldsymbol{y}), \tag{2.2}$$

It transforms to a penalized optimization problem when K^* is unknown:

$$\min_{\mathcal{T}} \{\mathcal{C}(\mathcal{T}, \boldsymbol{y}) + pen(\mathcal{T})\} \tag{2.3}$$

2.1.1. Parametric setting

In the parametric case, the detection depends heavily on what we are looking for in the signal $\boldsymbol{y}$. For example, searching for slope changes in a signal [7, 28] does not require the same modelling as detecting changes in the mean [31, 16]. We restrict our investigation to a cost function based on maximum likelihood. In this setting, the observations located in the k-th segment is supposed to be following a distribution Q depending on a set of parameters $\boldsymbol{\theta}_k$. More formally, we have that:

$$y_t \sim f(.; \boldsymbol{\theta}_k) \mathbb{1}_{t_k^*+1 \leq t \leq t_{k+1}^*},$$

with f being the density function of distribution Q. In other words, we suppose that all observations emanate from the same distribution Q but the values of $\boldsymbol{\theta}_k$ change abruptly at each change-point t_k^*. We supppose that $\boldsymbol{\theta} \in \Theta$ with Θ being a subset of $\mathbb{R}^d$ and compact. The cost function used to evaluate segments in this context is the negative log-likelihood. Hence, for a segment $y_{u:v}$ with $u < v$, we can write:

$$W(y_{u:v}) = -\sup_{\boldsymbol{\theta} \in \Theta} \sum_{i=u}^{v} \ln f(y_i; \boldsymbol{\theta})$$

In this case, the optimization problem 2.3 rewrites as:

$$\min_{\mathcal{T}=(t_0<t_1<\ldots<t_K<t_{K+1})} \{-\sup_{\boldsymbol{\theta} \in \Theta} \sum_{k=0}^{K} \sum_{i=t_k+1}^{t_{k+1}} \ln f(y_i; \boldsymbol{\theta}_k) + pen(\mathcal{T})\} \tag{2.4}$$

These models are very common in litterature and were adapted to various types of distributions Q. The investigations spread from the gaussian, to distribution belonging to the exponential family and even to discrete distribution.

2.1.2. Non-Parametric setting

The cost function for a segment can also be adapted for nonparametric statistical inference. Several strategies have been developed in the literature over time. These include the nonparametric maximum likelihood method [105, 24], kernel methods [36, 56], and rank-based methods [78, 97]. We will focus on the latter here for two main reasons. First, the ranked based methods were already known from the experts working with the Anses. Second, the methods we are presenting here can be adapted for left-censored observations.

Detecting a breakpoint in a signal can be done using a test statistic based on the ranks of the observations rather than their values. The rank of the ith observation is defined as $R_i = \sum_{j=1}^{n} \mathbb{1}(X_j < X_i)$. Moreover, we note $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i < t)$ the empirical cumulative distribution function (c.d.f.). The cost function is derived from the Wilcoxon/Mann-Whitney rank criterion. This is equivalent to running the test under the following assumptions:

- $\mathcal{H}_0$: there are no breaks in the $\boldsymbol{Y} = (Y_1, ..., Y_n)$
- $\mathcal{H}_1$: there is a change t^* such that $Y_1, ..., Y_{t^*}$ are distributed according $\mathbb{P}_1$ and $Y_{t^*+1}, ..., Y_n$ are distributed according to $\mathbb{P}_2$.

The rank statistic of the t-th observation is centered and is written as follows:

$$U_n(t) = \frac{2}{\sqrt{nt(n-t)}} \sum_{i=1}^{t} \left(\frac{n+1}{2} - R_i \right) \tag{2.5}$$

The test statistic for $\mathcal{H}_0$ and $\mathcal{H}_1$ is defined as:

$$S_n(t) = \hat{\Sigma}_n^{-1} U_n^2(t), \tag{2.6}$$

where $\hat{\Sigma}_n = \frac{4}{n}\sum_{i=1}^{n}(\hat{F}_n(X_i) - 1/2)^2$. Theorem 1 of [64] shows that under the null hypothesis the S_n are distributed according to a χ^2 distribution.
The non parametric test statistic was extended to multiple changepoint detection by [64]. The cost function W for a segment $y_{u:v}$ is defined as:

$$W(y_{u:v}) = -(v-u)\hat{\Sigma}_n^{-1}\overline{R}_{u:v}^2, \tag{2.7}$$

where $\overline{R}_{u:v} = \frac{1}{v-u}\sum_{i=u}^{v} R_i$ is the average rank of $y_{u:v}$.
In this case, the optimization problem 2.2 rewrites as:

$$\min_{|\mathcal{T}|=K} -\sum_{k=0}^{K} \sum_{i=t_k+1}^{t_{k+1}} (t_k + 1 - t_{k+1})\hat{\Sigma}_n^{-1}\overline{R}_{t_k+1:t_{k+1}}^2 \tag{2.8}$$

2.2. Searching for change points

Various methods for finding breakpoints have been described in the literature. They can be distinguished according to whether they provide an optimal solution to the problems 2.2 and 2.3 or an answer in the form of an approximation. Only the optimal methods are listed here. Approximation methods are not discussed, but there are plenty of them, such as sliding window methods [57, 61], bottom-up segmentation [18], and binary segmentation [99, 32]. Two methods that provide an optimal solution are presented below.

2.2.1. Optimal segmentation approach

The optimal segmentation alogirhtm brings an answer to problem 2.2. With a fixed K number of change-points, one can recursivly solve the optimization problem.

Algorithm 1 Optimal partition algorithm:

input : signal $y_{1:n}$, cost function $c()$, number of changepoints $K \geq 1$
Create C_1 a $n \times n$ empty matrix
for all (u, v) such that $1 \leq u < v \leq n$ **do**
  $C(u, v) \leftarrow c(y_{u:v})$
end for
if $K + 1 > 2$ **then**
  for $k = 2, ..., K$ **do**
    for all $u, v \in \{1, .., n\}$ such that $v - u > k$ **do**
      $C_k(u, v) \leftarrow \min_{u+k-1 \leq t < v} C_{k-1}(u, t) + C_1(t + 1, v)$
    end for
  end for
end if
$L \leftarrow (0, ..., 0)$ vector of size $K + 1$
$LK + 1 \leftarrow n$
$k \leftarrow K + 1$
while $k > 1$ **do**
  $s \leftarrow L(k)$
  $t^* \leftarrow \arg\min_{k-1 \leq t < s} C_{k-1}(1, t) + C_1(t + 1, s)$
  $L(k - 1) \leftarrow t^*$
  $k \leftarrow k - 1$
end while
Output: a list L of K estimated changepoints (with n as a last coordinate).

2.2.2. PELT algorithm

2.3. Estimation of the number of changes

Several methods are possible to select the best segmentation model. In the case of using a penalized criterion, this is equivalent of tuning the penalty term. This state of the art deals only deals with linear penalties, as mentioned in 2.1. Other types we will not discuss in this work (see for instance [35, 101, 95]). When no penalized criterion is used, there are other methods for selecting the best segmentation.

2.3.1. Different heuristics

Methods for penalty calibration

Methods using the optimal partitioning algorithm

2.3.2. Exploratory approach: CROPS algorithm

Algorithm 2 PELT algorithm

input : the data $y_1, ..., y_n$, the censoring threshold a, and the penalty term β_n

initialisations : $F(0) = \beta_n$, $R_1 = \{0\}$, $CP(0) = NULL$
for all $\tilde{t} = 1, ..., n$ **do** :

1. Compute $F(\tilde{t}) = \min_{t \in R_{\tilde{t}}} \{F(t) + W(y_{(t+1):\tilde{t}}, \hat{\lambda}_{(t+1):\tilde{t}}) + \beta_n\}$
2. Compute $\bar{t} = \arg\min_{t \in R_{\tilde{t}}} \{F(t) + W(y_{(t+1):\tilde{t}}, \hat{\lambda}_{(t+1):\tilde{t}}) + \beta_n\}$
3. Set $CP(\tilde{t}) = [CP(\bar{t}), \bar{t}]$
4. Set $R_{\tilde{t}+1} = \left\{ t \in R_{\tilde{t}} \cup \{\tilde{t}\} | F(t) + W(y_{(t+1):\tilde{t}}, \hat{\lambda}_{(t+1):\tilde{t}}) + \beta_n \leq F(\tilde{t}) \right\}$

end for

output : the vector of change-points CP.

Algorithm 3 CROPS algorithm

input : the data $y_1, ..., y_n$,
the bounds of the initial interval of penalties β_{min} and β_{max},
`PELT` algorithm
Compute $\texttt{PELT}(y_{1:n}, \beta_{min})$ and $\texttt{PELT}(y_{1:n}, \beta_{max})$
Define $\beta^* \leftarrow \{(\beta_{min}, \beta_{max})\}$ a list of vectors.
while $\beta^* \neq \emptyset$ **do**
  Define $(\beta_0, \beta_1) \leftarrow \beta^*(1)$
  if $m(\beta_0) > m(\beta_1) + 1$ **then**
    $\beta_{int} \leftarrow \frac{\mathcal{Q}_{m(\beta_1)}(y_{1:n}) - \mathcal{Q}_{m(\beta_0)}(y_{1:n})}{m(\beta_0) - m(\beta_1)}$
    $res \leftarrow \texttt{PELT}(y_{1:n}, \beta_{int})$
    From res store $m(\beta_{int})$
    if $m(\beta_{int}) \neq m(\beta_1)$ **then**
      $\beta^* \leftarrow \{\beta^*, (\beta_0, \beta_{int}), (\beta_{int}, \beta_1)\}$
    end if
  end if
  $\beta^* \leftarrow \beta^* \backslash (\beta_0, \beta_1)$
end while
output : Detailed segmentation for all $\beta \in [\beta_{min}, \beta_{max}]$.

2.4. Extension to the multivariate case

2.4.1. Fully multivariate detection

2.4.2. Subset multivariate detection

3. Chapter 3 : Problematic and data

Contents

3.1 General overview 19
3.1.1 Emission source 19
3.1.2 Substance diffusion 20
3.1.3 Impregnation levels and potential adverse effects 23
3.2 Environmental data specificity 24
3.2.1 Inherent data characteristics 25
3.2.2 Heterogeneity in distribution and sampling rates 26
3.2.3 Representation 27

Studies on environmental data befalls a multidisciplinary domain called environmental science that regroups various fields such as : physics, biology, chemistry, geography, ecology (and many more that won't be listed here). That important gathering of subjects induced an large collect of data coming from different source of information. As stated in [67], the emergence of environmental statistics comes from the obvious fact that much of what is learned on the environment is based on numerical data. Three broad types of areas of studies that we believe it is important to state in this work :

- **Baseline studies** aim at documenting the present knowledge and how environmental processes operate. Future changes will be define as any deviation from the standards identified by those studies.
- **Targeted studies** intend to characterize and assess the impact of planned or known changes (accidents, human activities).
- **Regular monitoring** is designed to detect patterns such as variations, trends or changes in important parameters.

In recent years, we can cite numerous applications that can be designated as environmental statistics and they span on a very large specter of problems [12, 37, 44, 104, 103, 87, 77, 70]. Phyto-pharmacovigilance is part of this type of study. It can be formally defined as a vigilance system that collects and analyzes monitoring data on phytopharmaceuticals (pesticides). The aim is to detect adverse effects associated with the use of these products as quickly as possible in order to protect the health of living organisms and ecosystems. This task falls under the responsibility of several agencies in Europe, including the European Food Safety Agency (EFSA), the European Chemical Agency (ECHA) and the European Environment Agency (EEA). In France, this task falls under the responsibility of the Agence nationale de sécurité sanitaire

de l'alimentation, de l'environnement et du travail (Anses) i.e. the French Agency for Food, Environmental and Occupational Health & Safety. Phytopharmacovigilance is an essential complement to the other tasks of the Agency, as one of its main tasks is to regulate by granting or denying authorizations for pesticide products[1].
Data collected or generated as part of phytopharmacovigilance allow to adjust, if necessary, the conditions of authorization for products currently placed on the market, e.g. by reducing the dosage, adjusting the conditions of use or withdrawing an authorization; to initiate management measures to reduce risk exposure, e.g. to protect people in the vicinity of treated areas; to contribute to the development of pre-market risk assessment methodologies used at the European level, if necessary.
The Anses does not collect the data it uses directly to accomplish its missions. It relies on its partner network to obtain data sets of interest. This partner network consists of other agencies (national or European), research laboratories, and associations. The decision to add a new source of information first requires a discussion on the coherence of the use of this source of information to provide an answer to the problem under study. For pharmacovigilance, the following information is of interest :

- Contamination of the environment - air, water, soil, food and drinking water - by residues including metabolites of pesticides.

- Exposure, impregnation and effects on living organisms and ecosystems as a whole: humans, livestock and wildlife, crops, flora, etc. Resistance phenomena in organisms targeted by these molecules: Pathogens, weeds, insects.

The purpose of this work is to assist agency experts in their task of monitoring pesticides use and effects on the territory. More precisely our problematic is the following : **How to locate in time and in space interesting signals from the data at disposal ?** It is essential to make an inventory and a brief description of the data that we will process for their study. To get a general idea of the information that the various data sets can provide, we propose the approach schematized in Figure 3.1. It is argued that if one knew and controlled all the numerical data in these three compartments and their relationship to each other, one could fully describe the lifetime of a chemical in its application. We divide this diagram into three compartments:

- the emitting source, which corresponds to the knowledge of the dose and the application methods of the considered substance This compartment corresponds to the description of the input data of a diffusion system in a medium.

- the diffusion of the substance, which depends on the ambient environment considered. combines information about the structure of the observed environment and the chemical properties of the substance. This compartment describes how the substance moves in a particular environmental medium.

- the impregnation and potential undesirable effects, which are simply the exposure and consequences of the substance's introduction into the environment.

[1] All their decision statements are available online `https://www.anses.fr/fr/decisions`.

In practice, we don't have a full control and knowledge of such information. We will see how to get partial information over this circuit in Figure 3.1. A general overview of the available information will be carried out in Section 1 The description of these data will allow us to identify the specific features and challenges that we will present in Section 2.

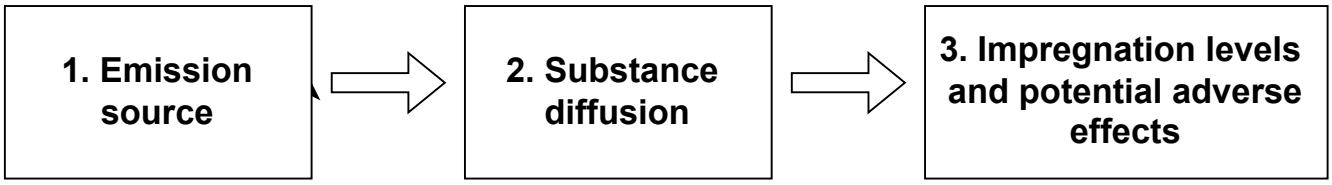

Figure 3.1: Three main compartments of a complete monitoring procedure.

3.1. General overview

This section illustrates the variety and scope of data sets that can be used for phytopharmacovigilance. We will show which sources of information we can assign to each of the subjects described in Figure 3.1. The inventory focuses on all data sets that can provide spatio-temporal information on the evolution of the substance during its emission, its dispersion, and the study of variations in potential adverse effects. We will illustrate several specific cases such as water, air, or soil quality monitoring. We will see that each context requires a selection of information that seems coherent and necessary. Wherever possible, graphical representations of examples will be shown. It is possible that some sources of information necessary for the analysis of certain phenomena are not available as open source. However, the producers of the data in question are systematically cited.

3.1.1. Emission source

The most accurate information about the sources of a substance's emissions could be obtained by quantified, parcel-based tracking of substance consumption. One could imagine a plot with one or more spatial points for each parcel. Each of these points could be linked to a temporal monitoring curve of the tonnage of a substance consumed in the part of the parcel it is located. Such a representation is not possible for several reasons. It is not legally possible to set up a monitoring system with sufficient resolution to identify a parcel owner. The anonymity of those involved in agriculture must be preserved in these investigations. In addition, such a system would only be feasible with considerable sampling effort. Nevertheless, three data sets can provide partial information on the source of pesticide emissions.
Specific pest species can be observed for each crop type. Mapping the crop types in an area can therefore provide a preliminary idea of the areas and periods of application of the substance being monitored. Some of this information is available in the graphical land register (GLR) [2]. This database corresponds to the application forms used by farmers to obtain financial aid

[2] Available at : `https://www.data.gouv.fr/fr/datasets/registre-parcellaire-graphique-rpg-contours-des-parce`

under the Common Agricultural Policy of the European Union (CAP). To be eligible for these grants, the crops grown on the plots must be declared. This dataset is a partial information, since the query of CAP is not mandatory. Therefore, the owners of the plots who have not applied for aid are not informed. Moreover, this register is renewed every year. It is possible that the information for certain parcels is not included in all annual editions of the GLR.
The use of a substance can also be indirectly seen in the sales data of crop protection products. The National Bank for the Sale of Pesticides by Authorized Distributors [3] (NBSD) lists and archives all such data. For the same reasons of anonymity, geographically fine resolution information is not available. The most accurate resolution corresponds to postal codes. It is the same for the temporal resolution that is not finer than the yearly resolution. Unlike CPL, this data set does not indicate the location and date of use of the substance. A purchaser may well be in a different location than the place of use of the substance they just purchased. Nevertheless, sales give a general indication of the intensity of use of a substance. A sudden increase in sales of a product may mean that its use is increasing in that area.
A final source of available information is cultural practice studies. They are conducted in the form of a questionnaire and are used to describe and characterize how farmers work on their land. These studies are very specific and focus only on certain types of crops when conducted. Three topics are addressed in the questionnaires. The first captures general information about the farm, such as commitment to a pesticide use reduction approach or related to agroecology. The second questionnaire is used to reconstruct the technical process on the plot. In other words, we examine the layout of the plot, its preceding crops or its irrigation. Finally, the use of plant protection products on the whole farm is investigated. The type and settings of the sprayer for the substance or the handling and protection of the user are criteria that are interrogated. In summary, cropping practice surveys provide much qualitative (rather than quantitative) information about the source of substance emissions. However, these surveys are conducted on an ad hoc basis and cover only certain types of crops. Only the results of the statistical departments of Agricultural ministry analyses [4] are available as open source, but not the raw data.

3.1.2. Substance diffusion

An ideal drug monitoring system would allow the concentration of the monitored molecule to be accurately quantified at any point in its network and at any time. Again, this is not possible in practice, most systems consist of a series of positioned stations at strategic places depending on the environmental medium investigated. Examples of monitoring in three different observing environments are presented here.
Surface water is water located on top of the Earth's surface, and is defined by opposition to underground waters. It is usually used specifically for terrestrial water bodies and littoral waters, the vast majority of which is produced by precipitation and runoff from nearby higher areas. The propagation of a substance is then assessed by a network of stations positioned on rivers or lakes sides and sampling directly from the water. All records are available on the

[3] Available at `https://geo.data.gouv.fr/fr/datasets/bdc2c6f21f70acccfea73445f68a5f0d6ee5b7c1`.
[4] An example can be found at :`https://agreste.agriculture.gouv.fr/agreste-web/disaron/Chd2009/detail/`.

site of the National Water Data and Repository Administration Service (Sandre[5] and Naïades sites[6]). A coherent geographical source of information to cross with the samples information is the hydrographic network of the area of study. This information is provided by the National Geographical Institute (IGN) [7]. This network allows to identify which stations are linked by a path of water and the flow direction identify an order between stations (which one is upstream/downstream). It will then provide a better understanding of the temporal dispersion of a substance in the rivers network. Figure 3.2 illustrates all stations of the Centre-Val de Loire region that made at least one sampled between 2007 and 2022 and the hydrographic regional network. Another complementary source of information is the hydro-ecoregions. Although the hydrographic network is the information with the most accurate geographic resolution, some parameters not informed by the IGN may influence the diffusion of a pesticide in water. Examples include climatic conditions and riverbed composition. Hydro-ecoregions (HER) are geographic units in which hydrographic ecosystems share common characteristics. The criteria by which they are delineated combine characteristics of geology, terrain, and climate [98]. The National Institute for Agriculture Food and Environment (INRAE) services provided such information[8]. The boundaries of these areas are shown in black in Figure 3.2.

A second example is air quality monitoring. Information on active substance concentrations is collected and available as open source on the website ATMO website[9]. This website aggregates all data from the regional air quality monitoring agencies (ASQAAs). This example illustrates the fact that the monitoring mission is highly dependent on the monitored environment. This change in propagation medium means that the coherent data sets are no longer the same (it seems fairly obvious that river information does not provide air quality information). Meteorological data is a relevant dataset for this application, especially any information that can be found about wind (wind direction, wind strength, etc.). Historical weather records are now available as open source on the Météo France website[10]. Figure 3.3 illustrates the cross-referencing of data from air quality monitoring stations with meteorological data.

The last example shows that environmental compartments are very different. The agency may also be interested in the quality of water for human use. Unlike the previous examples, the selection of consistent data sets for the study of this environmental medium is less obvious. Three main functions can be distinguished in a drinking water system [81]. The first corresponds to the sampling and collects all information about the raw water resource. The second one defines the production of the drinking water, i.e. its transport, storage and eventual treatment. The distribution to the user is the last function. In order to obtain information that allows the observation of these three functions, the required data sets must be diverse and of different nature. The sampling part is observed using information on the quality of the raw water and therefore requires data sets on environmental quality (as in the previous two examples). The other two functions, on the other hand, are related to human activities and therefore require

[5] https://www.sandre.eaufrance.fr/

[6] https://naiades.eaufrance.fr/donnees-disponibles

[7] See :https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo

[8] The open data HER shapefiles : https://geo.data.gouv.fr/fr/datasets/1135b7fd3ca5de69a080f04c31cdc1216a9a34e0

[9] https://www.atmo-france.org/

[10] https://public.opendatasoft.com/explore/dataset/donnees-synop-essentielles-omm/table/?flg=fr&sort=date

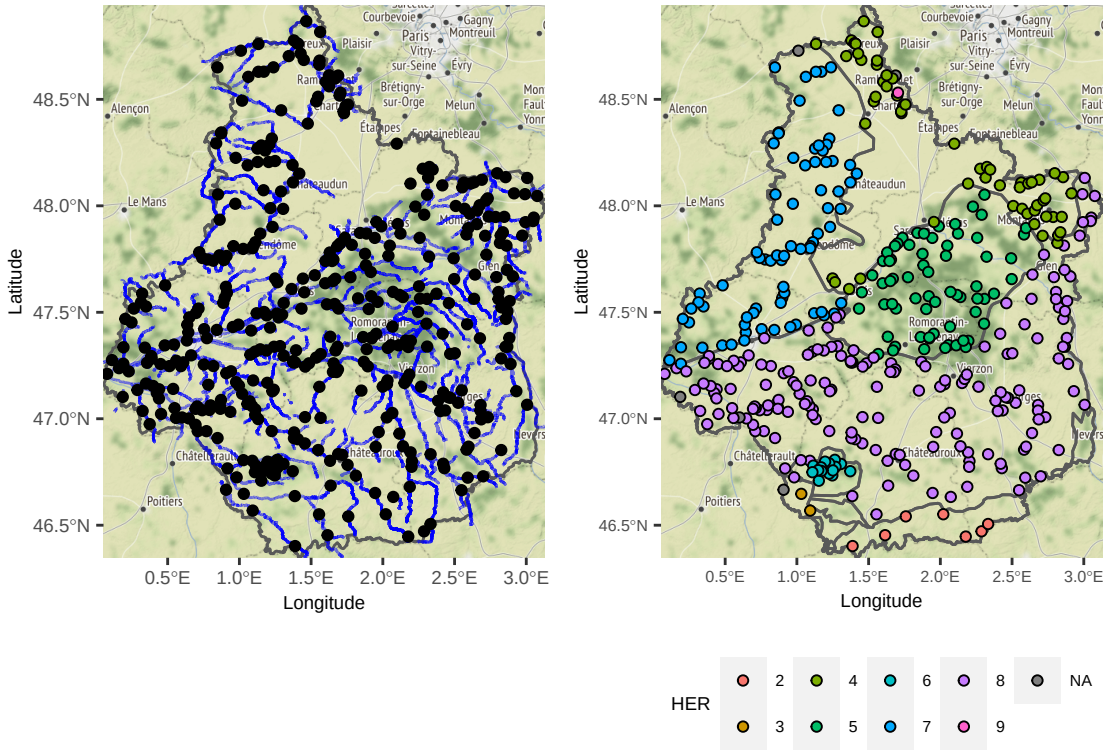

Figure 3.2: Stations monitoring water suface in the Centre-Val de Loire french region. Two different geographical scales are represented. The underlying hydrographic network linking all stations is plotted on the left, the stations are colored according to their hydroecoregion on the right.

a completely different spatial approach. For example, observation at the administrative level (e.g., departmental level) is consistent. If finer resolution is desired, we can consider water management units (WMUs) or other scales described in [81]. All the data is provided by the Department of Health[11].

Thus, it is clear that each environmental compartment is examined using a variety of data sets. The list in this section is not exhaustive. We can add to it the study of soil quality, food production and distribution, animal feed, groundwater, indoor air and dust... In particular, we note that the spatial analysis of the dispersion of a substance in an environment depends very much on the observed network and the mechanisms that govern that environment. For example, the hydrographic network of a region depends both on the natural geography of the area and on human activity (presence of irrigation canals in the network), while the distribution units of water intended for human consumption are entirely determined by human activity.

[11] Open source data at: `https://www.data.gouv.fr/fr/datasets/resultats-du-controle-sanitaire-de-leau-du`

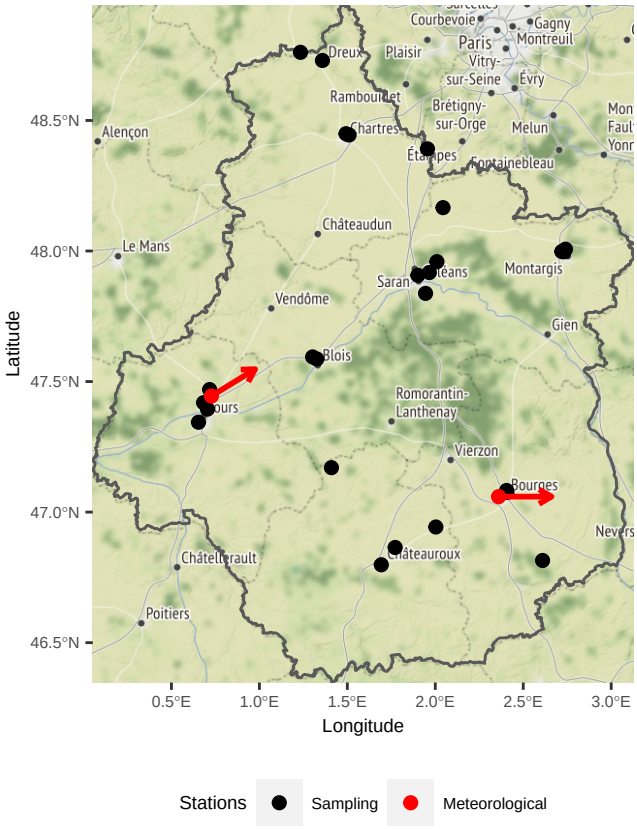

Figure 3.3: All active stations measuring air quality on March the 1st of 2021 coupled with meteorological stations active that day. The main wind direction and speed measured that day is mapped with the red arrows.

3.1.3. Impregnation levels and potential adverse effects

Once a pesticide has been applied and has spread in an environmental compartment, it is interesting to observe how ecosystems are affected by the active substance. Two aspects can be distinguished: the exposure to which ecosystems are subjected, i.e., their impregnation, and the possible adverse effects that result. As is often the case in statistics [88], it is complicated to identify direct relationships between cause and effect. Rather, these two aspects are linked through correlational relationships. The data sets are not of the same type. Several examples of datasets illustrating those two aspects are given here. Very few data are available as open source, as they all fall under the General Data Protection Regulation (GDPR). We therefore limit ourselves to mentioning the names of the actors responsible for the collection of the data and to giving a general overview of the content of the data.

Concerning the impregnation of the environment, the main issue here is measuring the concentration of active ingredients in the living organisms that make up the exposed ecosystems. For example, the Anses, in collaboration with Santé Publique France, will be conducting a

new study called Pestiriv [12] on this topic in the coming months. In France, a large part of the rural population lives in wine-growing areas, which explains the importance of this project. As part of this study, hair and urine samples will be taken from people living in wine-growing areas. This will allow the exposure of farmers to pesticides to be measured. The deposition of pesticides in bee matrices is another example of impregnation. The Technical and Scientific Institute of Apiculture and Pollination (ITSAP) is the institute responsible for monitoring this environmental medium. We look for pesticide concentrations in pollen in beehives..
The databases for monitoring potential adverse events are medical registries. There are human and animal health databases. For human health, the Phytattitude network was developed by the Mutual Agricultural Health Insurers (MSAs). It is a network where any professional who comes into contact with phytosanitary products can indicate if he/she has health problems. This organization collects data through spontaneous reports from agricultural actors or during scheduled visits by nurses or doctors. Another source is the medical-administrative databases of the MSA. They collect information on farmers' health care reimbursements. Second, poison control centers are involved in adverse effect surveillance. They provide toxicovigilance information on toxicovigilance for the entire population. Much information about acute health problems comes up through these information channels. For chronic health problems, there is the National network of vigilance and prevention of professional pathologies (RNV3P), whose role is to identify emerging or re-emerging occupational health risks. Finally, the AGRICAN cohort (AGRIculture and CANcer) of the François Baclesse Center is used to measure the health status of the agricultural population compared to the general population (especially in terms of cancer burden). It is therefore also part of the Anses partner network of partners. Regarding animal health, INRAE provides a database on veterinary toxicovigilance (GIS Toxinelle), and the Biodiversity French Office (OFB) on wildlife toxicovigilance of wildlife. The Department of Agriculture provides additional information, such as acute mortality in bees, and its 500 ENI biovigilance program is also part of the available databases. This is a program to monitor the impact of agricultural practices on biodiversity.
This section explains the diversity of the actors that are involved in a partnership with the Anses and the extent of the number of data sets that can be of use in a monitoring mission. It also highlights the need for regular discussion with experts to identify and properly manipulate this information. In this work, we made a selection of datasets focused on surface water monitoring. This allowed us to narrow down the many possible research directions and frame the work done in this thesis. Figure 3.4 summarizes the datasets that were selected to develop an algorithm for detecting anomalous spatiotemporal signals in surface waters. This scheme is presented in section 5. Impregnation and adverse effects data are not available, so this aspect of the monitoring scheme is not addressed.

3.2. Environmental data specificity

Although the sources of information used to develop models vary, common features can be identified. Three aspects of different nature are presented in this section. All illustrations are

[12] see full description here :https://www.anses.fr/fr/content/lancement-de-pestiriv-une-%C3%A9tude-in%C3%A9dite-sur-l%E2%80%99exposition-aux-pesticides-des-personnes-vivant

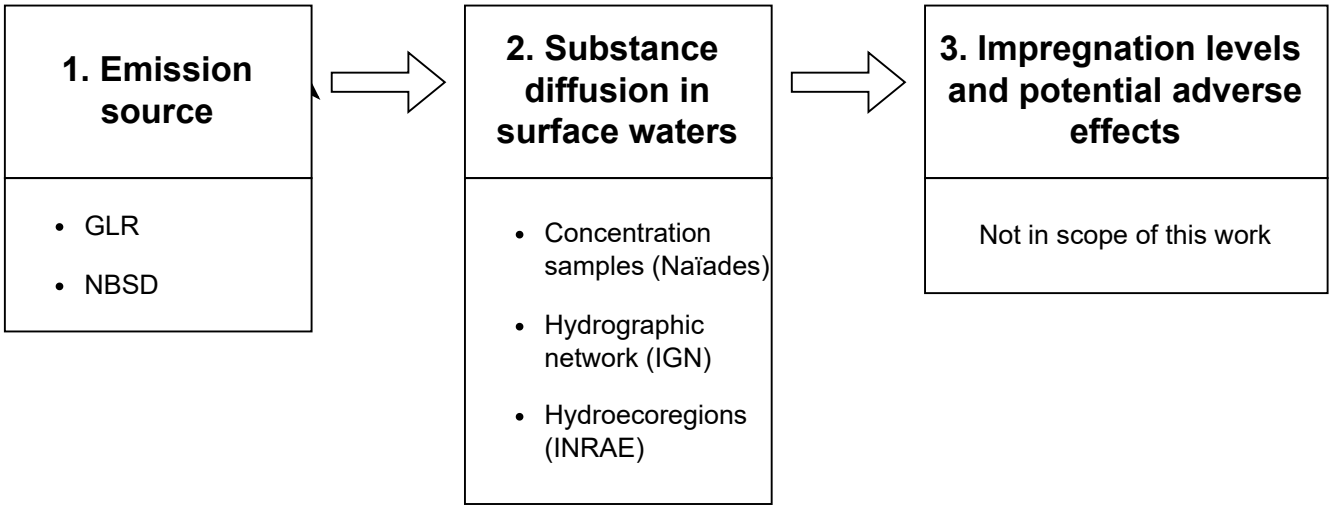

Figure 3.4: Data sets selected in this work.

made with data of prosulfocarbe concentration in the Centre-Val de Loire french region.

3.2.1. Inherent data characteristics

The first specificity of concentration measurement arises from the problem of measuring a chemical substance in a sample. In applied chemistry, any measuring device is characterized by two types of limits.

- The detection limit (LOD): This is the smallest concentration value in a sample that can be distinguished from zero with certainty.
- The limit of quantification (LOQ): This is the smallest concentration value of a substance in a sample that can be measured with certainty.

It is important to know that these two limits are determined by the instrument that performs the measurement. It happens that geographical areas are covered by stations that do not have the same equipment. In this case, there are several LOQ values within the samples taken in that area. In the case of surface waters, for example, the contracts for the selection of monitoring laboratories are awarded by the water agencies. These agencies, six in number, cover an area larger than the French administrative regions. This means that if the scale of an administrative region is taken as the basis for a study, this region may fall under the jurisdiction of two different water agencies and therefore the measuring instruments may be different. Moreover, the same station may change its measuring equipment over time. Station equipment contracts are renewed periodically, but renewal does not guarantee that the same equipment will be maintained. All those caracteristics are illustrated in Figure 3.5. These two limits of accuracy mean that concentration data are left-censored. However, unlike the classical censoring problems described in the literature, the values of the censoring thresholds are known. Several methods have been developed to handle this type of data. We can cite the imputation

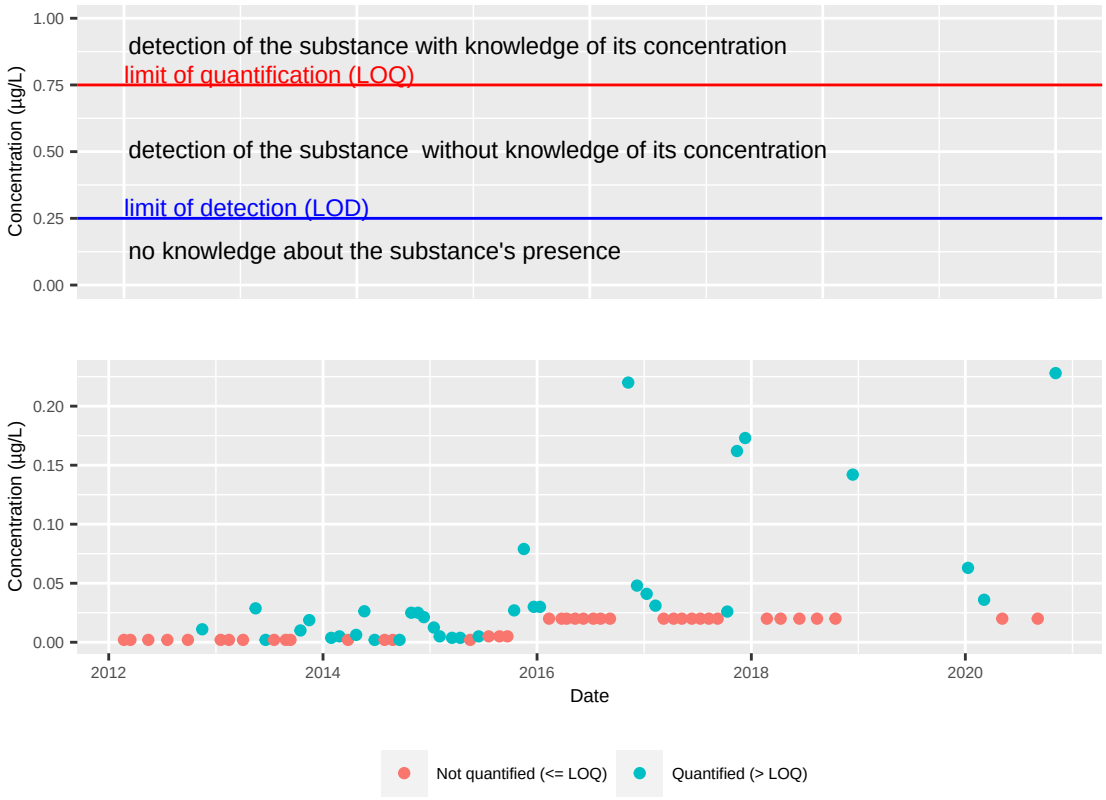

Figure 3.5: Censorship illustration. The top Figure sums up the limits of measurements effects. The bottom Figure shows the consequences of censorship on the samples of a station located in the Centre-Val de Loire region. This station changed its equipment in 2016, the LOQ change values.

of values to replace the LOQ values present in the set of concentration values, the use of the maximum likelihood estimator or the Kaplan-Meier estimator (see [33, 22]). In Chapter 4, we will show which method was chosen to handle this type of data in this thesis.

The concentration data also show strong tails in their distribution. In Figure 3.6 there are not only many low concentration values (corresponding to the measured values that did not exceed the quantification threshold), but also high concentration values. The maximum value is almost 3 µg/L. The data used in Figure 3.6 are the measured values from a period when the substance in question was typically used. This ensures that we have a high quantification rate and thus many fully observed samples. The concentration data can therefore be summarized as left-censored heavy-tailed data.

3.2.2. Heterogeneity in distribution and sampling rates

The second feature of these data is the spatial and temporal heterogeneity of the measurements. This is not the case for all pesticide monitoring data. This makes it a particularity of the

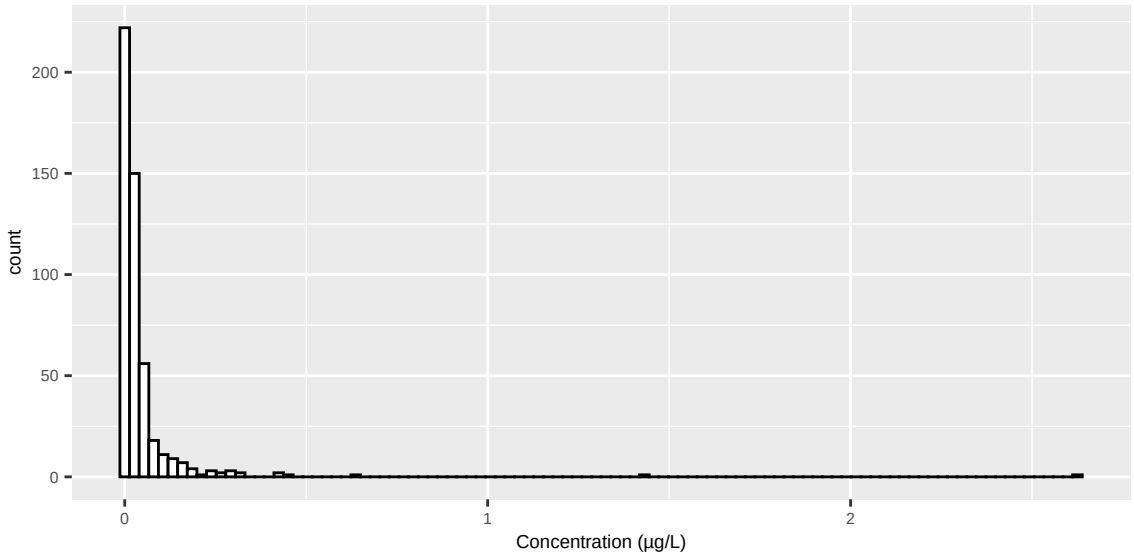

Figure 3.6: Histogram of all measurements of prosulfocarbe made between 2017-10-01 and 2018-02-01 in the Centre-Val de Loire region.

French network. Figure 3.5 already shows that the rhythm of the measurements for the same station is irregular in time. Figure 3.7 shows that for two spatially neighbouring stations the measured values are not synchronous in time. It can also be seen that the stations take very few measurements and do not take the same number of measurements. In Figure 3.7 we see that it is not possible to compare the measurements of the two existing stations. The stations literally sampled in non-overlapping time periods. Therefore, to derive information from these data, one must work on a different scale than that of the station. Aggregating the data from the two stations results in a time series that is more evenly sampled over time. Thus, there is a trade-off between the number of data available to make a statistical statement about a spatial area and the accuracy of that spatial extent.
Figure 3.8 illustrates that alongside heterogeneity in the stations sampling rhythms, spatio temporal data are not distributed in a homogeneous way over the territory. Depending on which region of the area of study the samples originate, the underlying distribution differ. Thus, the data are heterogeneous in space. This Figure also illustrates that the distribution of concentration values can drastically change over time. Looking at the samples of station 03189000, there is a break point just before the year 2015 in the station concentration values. The same can be saif for the year 2016 in Figure 3.5. Thus the data are heterogeneous in time. The spatio-temporal heterogeneity affects the sampling rhythms and the distribution of the concentrations.

3.2.3. Representation

The final problem associated with spatiotemporal data is their representation. The minimum number of dimensions to describe them is three: two for space as a pair of longitudes and latitudes; one for time. [21] presents in his book several methods of descriptive statistics for

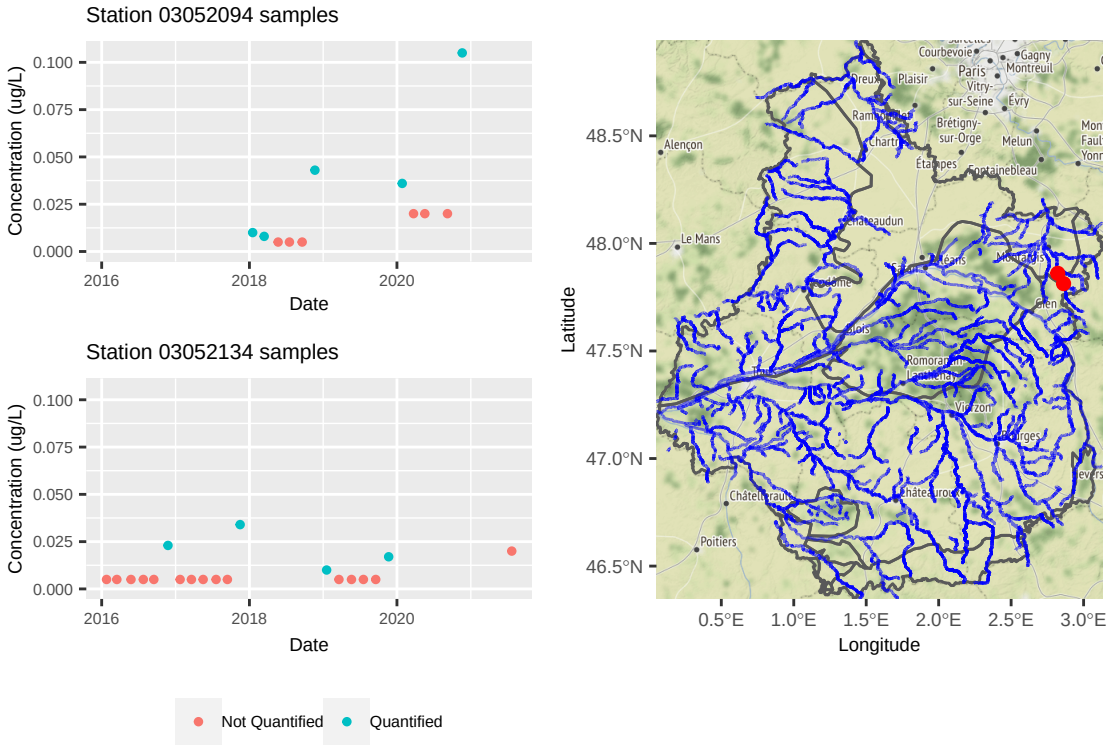

Figure 3.7: Spatial and temporal heterogeneity in sampling. The Figures on the left represent all the samples of two neighbouring stations. The map on the right shows the position of those stations.

the representation of such data. The most intuitive method to plot spatiotemporal data are marginal and conditional plot. There are several ways to make them.
The first one are **time-series plot**. The figures presented so far in this section are one possible example of this representation where one commits to a geographic position for observing a time series. Figure 3.9 displays the daily maximal concentrations of prosulfocarbe in three adjacent hydroecoregions (report to Figure 3.2 for spatial representations of the HER). Clear correlation is observed between the time series. However, the signal of HER 4 doesn't seem to be exactly the same and shows earlier signs of a seasonal pattern and of higher quantified values. Another example is shown in Figures 3.8 and 3.7 that allowed to see the spatial heterogeneity of sampling rates and distribution. The problem is that we cannot get a complete view of the spatial distribution of the data. The stations compared in both figures were hand-selected, and no information is available about the others. By selecting the HER scale, we insert expert information in the visualisation. Note that each choice has its advantages and its drawbacks. For instance, if looking at HER 4 in Figure 3.2, stations that belong to it are separated in three clear clusters (one in the north, the other in the east and last one in the south west of the HER). As though they belong to the same HER, grouping the data of these stations is not necessarily an ideal choice as they are still far away..

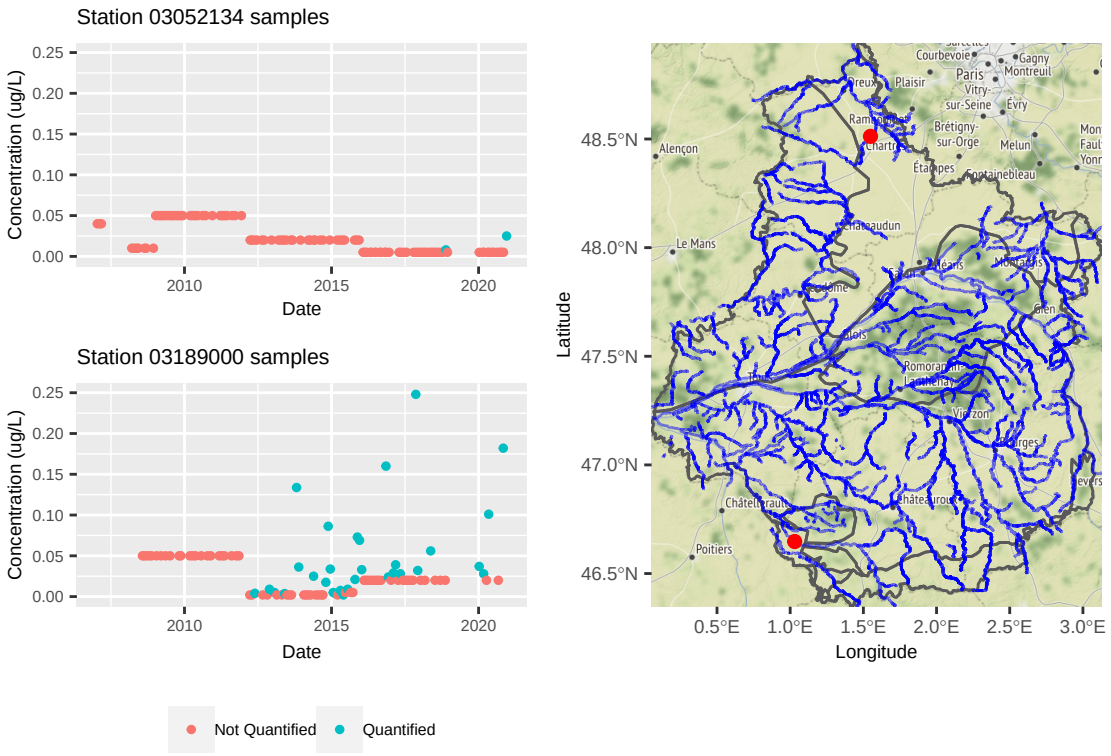

Figure 3.8: Spatial and temporal heterogeneity in distribution. The Figures on the left represent all the samples of two stations. The map on the right shows the position of those stations.

Second, the **spatial maps plot** can also be an effective way of extracting information from these data. It consists in a temporal sequence of maps as in Figure 3.10. A clear seasonal pattern can be identified from this Figure. Two problems emerge from this representation method. First, it takes an extensive number of maps to exhibit significant patterns. Secondly, the four seasons of the year were chosen to make a temporal segmentation of the two observations year. We introduced external knowledge to our analysis, Same as previously, even though it is a coherent choice, it has it s drawbacks. For instance, it cannot take into account the years where treatment started earlier or later due to climatic conditions. Furthermore, it can't help to determine precisely the nature of what is temporally changing in the signal. only the summarised indicator of quantification rate is used. Nothing is known about the maximum or the mean concentrations. Plotting all the maps displaying other indicators would be a tedious task.

Another example of representation is the **space (1-D)/time plots** shown in Figure 3.11. It consists in fixing a dimension in space whether longitude or latitude and to plot some indicators summarised on that dimension against time. In Figure 3.11, the longitude was cut into regular intervals and time was segmented in seasons. The plot shows once again a very clear temporal seasons that appears in time in the region. The location of those changes are not clearly displayed by those plots though. We are not able to determine if the large quantification rates

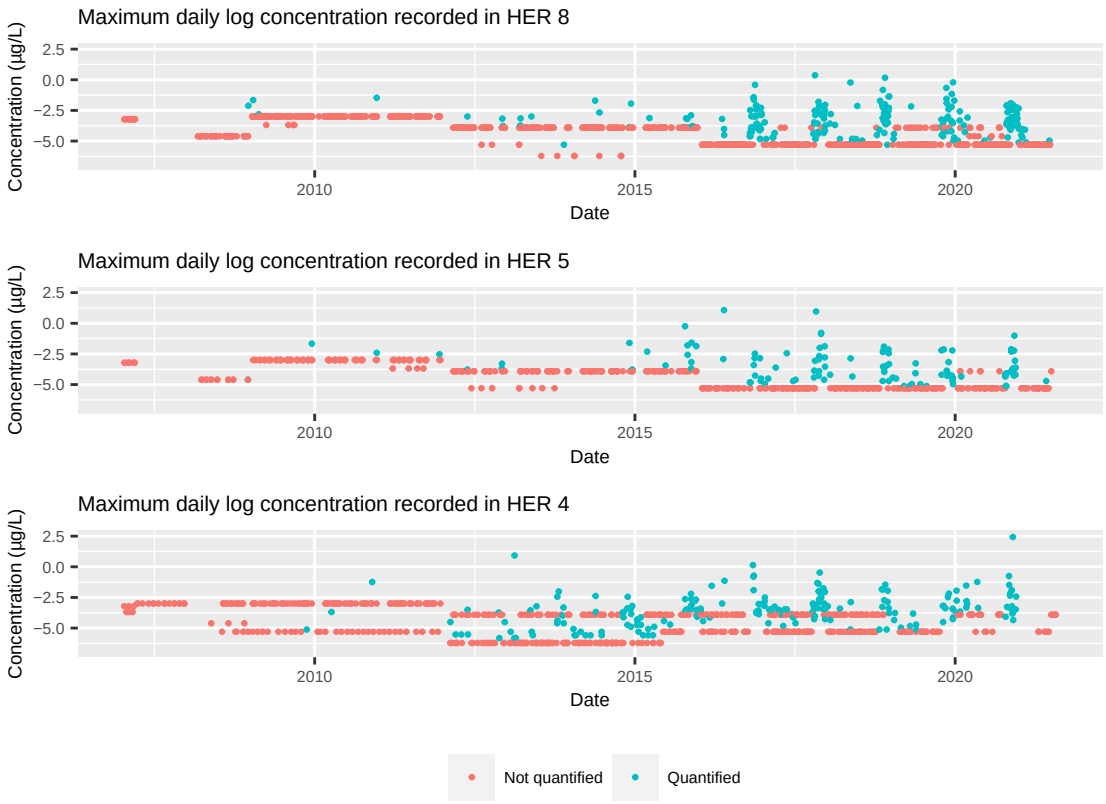

Figure 3.9: Time series plot of HER 8,5 and 4 daily maximum concentrations. The log scale was used for a easier visualization.

are in the central areas of the region or not (given that we miss the latitude information). Although those plots provide an excellent first overview and can help to quickly understand the structure under spatiotemporal data, it does not provide sufficient precision to localize in space and time some interesting signals.
All of those representation methods can't capture all the information carried by the data set. We propose to handle it in a dynamic way using the application environment R-shiny. Integrating dynamic response to a user's requests is an efficient way to combine a descriptive view with models results. The application overview will be given in section 6.

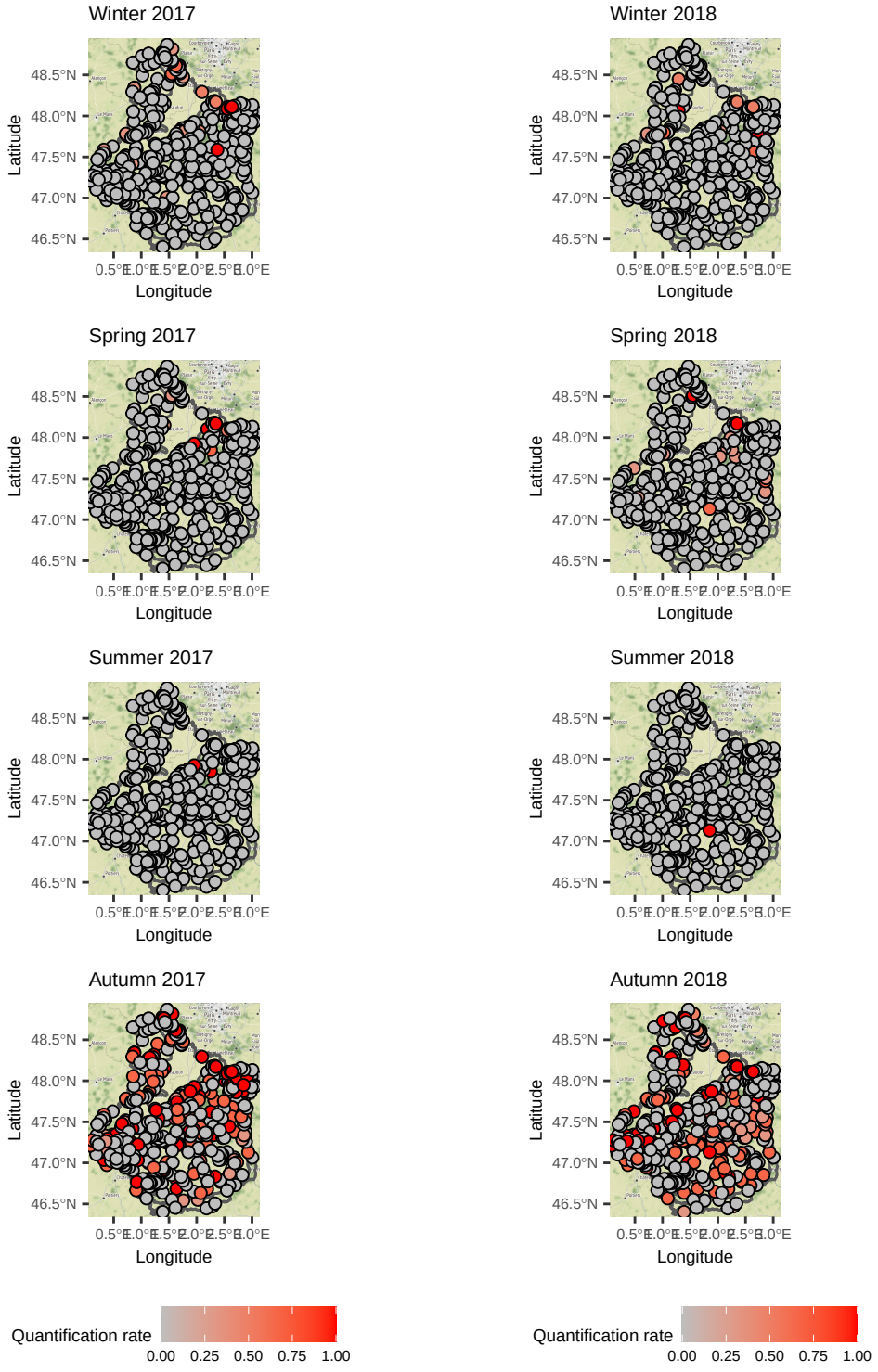

Figure 3.10: Spatial maps in time. The prosulfocarbe's quantification rate of each station was computed for each season of 2017 and 2018.

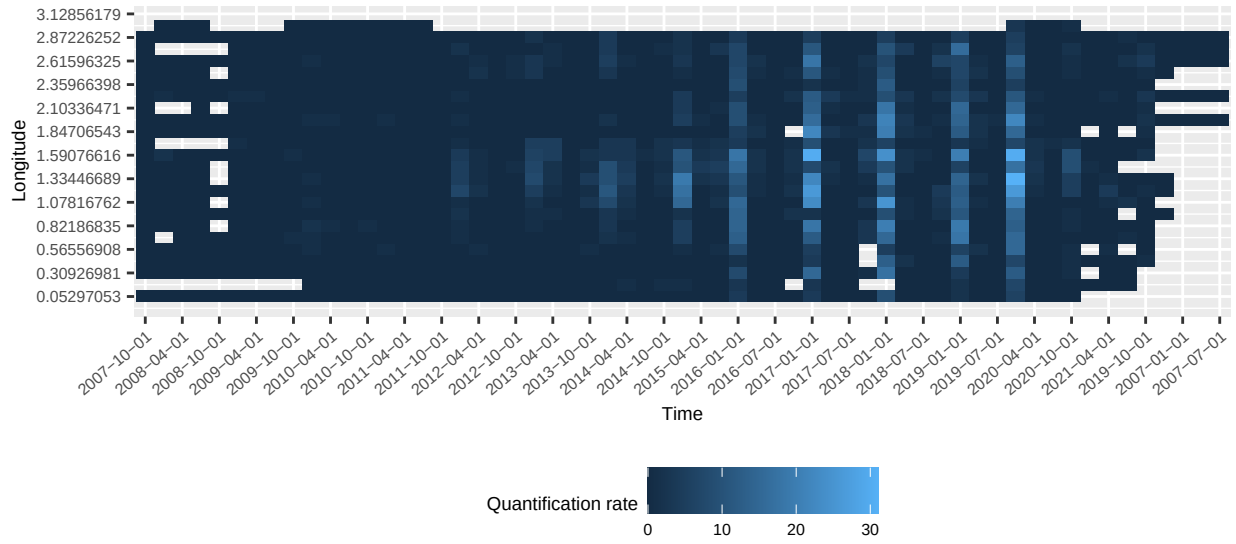

Figure 3.11: Space (1-D)/Time plots. Grey tiles correspond to location and moment were no sample were collected.

4. Chapter 4 : Change point detection in left censored observations

Contents

4.1 Model 33
4.1.1 Estimating change-points locations with $K^\star$ known 34
4.1.2 Estimating change-points locations with $K^\star$ unknown 36
4.2 Estimation procedure 37
4.2.1 Estimating the λ's 37
4.2.2 Estimation procedure for $K^\star$, $\boldsymbol{t}^\star$ and $\boldsymbol{\lambda}^\star$ 37
4.3 Simulation study 38
4.3.1 Tuning hyper-parameters of the Newton-Raphson method 39
4.3.2 Tuning the parameters of the change-point detection method 41

Faced with the singular characteristics of the data presented in Chapter 3, it is interesting to set up methods for an automatic detection of relevant signals in order to support expert analysis. Left censored data are characteristic of concentration records due to the inherent limits of qualification (LOQ) in the sample analysis. In this chapter, a new framework is proposed in order to detect changepoints in left censored signals. The model used to describe the data involves a left censored Weibull distribution. A changepoint detection method is derived from such modeling. We propose the pruned exact linear time (PELT) search method to do so. We will show that theoritical assumptions are met in the context of left censored data for both the convergence of our estimator and the use of the PELT algorithm. Experiments on simulated datasets are carried out, and the proposed approach is compared with related method. The issue of the LOD and of the LOQ has been addressed here by supposing that the data was left-censored, and that the censoring threshold, the LOQ, is a known constant. This approach is rather common in concentration monitoring studies, and is referred to as an upper bound configuration [8]. To our knowledge, using changepoint detection on left-censored distributions for analysing pollutant concentration represents a new contribution to the field.

4.1. Model

In the following, let us suppose that one has a series of observed data $y_1, \ldots, y_n$, which is the outcome of a random vector $Y_1, \ldots, Y_n$. The variables Y_i are recorded sequentially, although the recording times are not necessarily equidistant. Thus, the indices in Y_i are only indicators of the order of appearance in the sample, and not of the observation times. Furthermore, Y_i

are supposed to be independent. We are interested here in left-censored Weibull distributions, where the censoring threshold in known and fixed, and where potential changes in the scale parameter of the Weibull distributions may occur independently. In this chapter, we will suppose that the shape parameter σ is known and invariant in time. The case where this parameter is unknown is treated in Chapter 5. The definition of the true model writes as follows:

Definition 4.1.1. *Suppose that $Y_1, ..., Y_n$ are independent random variables, such that Y_i is a left-censored Weibull distribution, depending on a censoring threshold $a > 0$ (a is supposed to be known and fixed throughout) and on some shape parameter σ and a scale parameter $\lambda_k^\star \in \Theta \subset]0, \infty[$, for $t_{k-1}^\star \leq i \leq t_k^\star$, $k = 1, ..., K^\star$. The c.d.f. of Y_i is*

$$F_{Y_i}(x) = F(x; \lambda_k^\star, \sigma) = \left(1 - e^{-(\lambda_k^\star x)^\sigma}\right) \mathbb{1}_{\{x \geq a\}} \ , \tag{4.1}$$

The vector $\boldsymbol{t}^\star = (0 = t_0^\star < t_1^\star < ... < t_{K^\star-1}^\star < t_{K^\star-1}^\star = n)$ is the vector containing the change-points, and $\boldsymbol{\lambda}^\star = (\lambda_1^\star, ..., \lambda_{K^\star}^\star)$ is the vector of parameters associated to the $K^\star$ regimes.

In general, $K^\star$, $\boldsymbol{t}^\star$ and $\boldsymbol{\lambda}^\star$ are all unknown and one would estimate them starting from the data $y_1, ..., y_n$. Estimation will be carried out by minimising a contrast function (can also be desiganted by the term cost function), which would be the -log-likelihood in this framework. In the subsequent, we will consider two separate cases, according to whether the number of regimes $K^\star$ is priorly known or not.

4.1.1. Estimating change-points locations with $K^\star$ known

First, we shall consider the case where the number of regimes is known, and one looks for the estimates of $t_k^\star$, $k = 1, ..., K^\star - 1$ and $\lambda_1^\star, ..., \lambda_{K^\star}^\star$. We define the following criterion, based on the negative log-likelihood of the observed sample :

$$\mathcal{C}_{Y_{1:n}}(\boldsymbol{t}, \boldsymbol{\lambda}) = \sum_{k=0}^{K^\star-1} W(Y_{(t_k+1):t_{k+1}}, \lambda_k) = \sum_{k=0}^{K^\star-1} \sum_{i=t_k+1}^{t_{k+1}} -\ln f(Y_i; \lambda_k, \sigma), \tag{4.2}$$

where $Y_{u:v} = (Y_u, ..., Y_v)$, $\boldsymbol{t} = (0 = t_0 < t_1 < ... < t_{K^\star} = n)$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_{K^\star}) \in \Theta^{K^\star}$, and

$$f(Y_i; \lambda_k, \sigma) = \left(1 - e^{-(\lambda_k a)^\sigma}\right)^{\mathbb{1}_{\{Y_i = a\}}} \left(\sigma \lambda_k (\lambda_k Y_i)^{\sigma-1} e^{-(\lambda_k Y_i)^\sigma}\right)^{\mathbb{1}_{\{Y_i > a\}}} . \tag{4.3}$$

If one denotes

$$\mathcal{T}_{K^\star} = \{\boldsymbol{t} = (t_0, ..., t_{K^\star}), 0 = t_0 < t_1 < ... < t_{K^\star-1} < t_{K^\star} = n\}, \tag{4.4}$$

the set of all possible partitions of $Y_1, ..., Y_n$ into $K^\star$ regimes, then, the maximum likelihood estimates of $\boldsymbol{t}^\star$ and $\boldsymbol{\lambda}^\star$ are defined as the quantities realising

$$(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\lambda}}) = \arg \min_{\boldsymbol{t} \in \mathcal{T}_{K^\star}, \boldsymbol{\lambda} \in \Theta^{K^\star}} \mathcal{C}_{Y_{1:n}}(\boldsymbol{t}, \boldsymbol{\lambda}) \tag{4.5}$$

In the following, we'll consider a slightly different writing of this optimisation problem. Let us define the maximum likelihood estimate in the segment $Y_{(t_k+1):t_{k+1}}$ by :

$$\tilde{\lambda}_k = \arg\min_{\lambda\in\Theta} \sum_{i=t_k+1}^{t_{k+1}} -\ln f(Y_i;\lambda_k,\sigma) \ , \tag{4.6}$$

where

$$\sum_{i=t_k+1}^{t_{k+1}} \ln f(Y_i;\lambda_k,\sigma) = \sum_{i=t_k+1}^{t_{k+1}} \ln\left(1-e^{-(\lambda_k a)^\sigma}\right)\mathbb{1}_{\{Y_i=a\}} + \sum_{i=t_k+1}^{t_{k+1}} \Big(\ln(\sigma\lambda_k)+(\sigma-1)\ln(\lambda_k Y_i)-(\lambda_k Y_i)^\sigma\Big)\mathbb{1}_{\{Y_i>a\}}, \tag{4.7}$$

In this case, by plugging $\tilde{\lambda}_k$ in Eq. 4.5, the initial optimisation problem is reduced to computing $\hat{\boldsymbol{t}}$,

$$\hat{\boldsymbol{t}} = \arg\min_{\boldsymbol{t}\in\mathcal{T}_{K^\star}} \mathcal{C}_{Y_{1:n}}(\boldsymbol{t},\tilde{\boldsymbol{\lambda}}) = \arg\min_{\boldsymbol{t}\in\mathcal{T}_{K^\star}} \sum_{k=0}^{K^\star-1} W(Y_{(t_k+1):t_{k+1}},\tilde{\lambda}_k) \tag{4.8}$$

Several remarks may be made at this point. First, one should notice that Eq. 4.6 may not be explicitly solved. Nevertheless, one may check that second derivative with respect to λ_k is strictly positive, ensuring the existence of a minimum. The illustration is provided in Appendix A. From a practical point of view, in the numerical implementation to be presented in the next sections, a Newton-Raphson approach will be used for computing $\tilde{\lambda}_k$. Second, the search for $\hat{\boldsymbol{t}}$ in Eq. 4.8 would normally require a $\mathcal{O}(n^{K^\star-1})$, but we will reduce it to at most a quadratic one, using dynamical programming and other heuristics.
The estimate $\hat{\boldsymbol{t}},\hat{\boldsymbol{\lambda}}$ may be shown to have consistency properties. In order to establish them, some further notations and hypotheses are being needed.

(H1) Θ is compact and there exists $\Delta^\star_{\boldsymbol{\lambda}} > 0$ such that $|\lambda^\star_k - \lambda^\star_{k-1}| > \Delta^\star_{\boldsymbol{\lambda}}$, for all $k = 2,...,K^\star$.

(H2) If one denotes $\tau^\star_k = \frac{t^\star_k}{n}$, $k = 0,...,K^\star$, the normalized configuration, then $\tau^\star_k$ is constant when the sample size n varies, for all $k = 0,...,K^\star$.

(H3) There exists $\Delta^\star_{\boldsymbol{\tau}} > 0$ such that $|\tau^\star_k - \tau^\star_{k-1}| > \Delta^\star_{\boldsymbol{\tau}}$, for all $k = 1,...,K^\star$.

The first hypothesis mainly aims at ensuring sufficient conditions for the identifiability of the model, by imposing a minimum gap between two consecutive λ's. The second hypothesis, which is also the strongest one, implies that change-point locations are independent of the scale and frequency at which the data is sampled. This hypothesis will also allow us to derive the asymptotic behaviour of the estimate, when the sample size is sufficiently large. One should note here that a larger sample means a finer scale for sampling the data and not an extension of the period of observation. Eventually, the third hypothesis checks that each regime contains sufficient data for obtaining reliable estimates for the λ's.
With the above notations and definitions, one may state the following result:

Proposition 4.1.1. *Under the hypotheses (H1)-(H3), the maximum likelihood estimate is weakly consistent*

$$(\hat{\boldsymbol{\tau}},\hat{\boldsymbol{\lambda}}) \xrightarrow[n\to\infty]{\mathbb{P}^\star} (\boldsymbol{\tau}^\star,\boldsymbol{\lambda}^\star) \ , \tag{4.9}$$

where $\hat{\boldsymbol{t}}$ *and* $\hat{\boldsymbol{\lambda}}$ *are computed by solving Eq. 4.8 and 4.6, and* $\hat{\boldsymbol{\tau}} = \frac{\hat{\boldsymbol{t}}}{n}$ *is the estimate of the normalized configuration.*

This proposition is a particular case of Theorem 2.2 in [53]. A detailed proof checking that hypotheses (H1)-(H3) are sufficient is provided in Appendix A.

Not only is the maximum likelihood consistent, but one may equally derive its consistency rate:

Proposition 4.1.2. *Under the hypotheses (H1)-(H3),* $\{n\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^\star\|_\infty\}$ *and* $\{\sqrt{n}\|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star\|_\infty\}$ *are uniformly tight in probability:*

$$\begin{aligned} \lim_{\delta\to\infty}\lim_{n\to\infty} \mathbb{P}^\star\left(n\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^\star\|_\infty \geq \delta\right) = 0 \\ \lim_{\eta\to\infty}\lim_{n\to\infty} \mathbb{P}^\star\left(\sqrt{n}\|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star\|_\infty \geq \eta\right) = 0. \end{aligned} \tag{4.10}$$

4.1.2. Estimating change-points locations with $K^\star$ unknown

In practice, $K^\star$ will be most often unknown, and one will need to estimate it also. In this case, one may define a penalised criterion,

$$\tilde{\mathcal{C}}_{Y_{1:n}}(K, \boldsymbol{t}, \boldsymbol{\lambda}) = \sum_{k=0}^{K^\star-1} W(Y_{(t_k+1):t_{k+1}}, \lambda_k) + \beta_n K \ , \tag{4.11}$$

where β_n is a positive sequence.

We follow the same approach as in [53] to prove that under some further usual assumptions on β_n, the estimates obtained by minimising the penalised criterion in Equation 4.11 are consistent. First of all, let us assume that $K^\star$ is bounded from above, that is there exists some $K_{\max} \in \mathbb{N}$, such that $K^\star \leq K_{\max}$. Equivalently, one may suppose that there exists a minorant of $\Delta^\star_{\boldsymbol{\tau}}$:

(H4) There exists $\Delta_{\boldsymbol{\tau}} > 0$, such that $\Delta^\star_{\boldsymbol{\tau}} > \Delta_{\boldsymbol{\tau}}$. In this case, the maximum number of regimes may be written as $K_{\max} = \frac{1}{\Delta_{\boldsymbol{\tau}}}$.

For each $K = 1, ..., K_{\max}$, define the set of possible configurations:

$$\mathcal{T}^\Delta_K = \{\boldsymbol{\tau} = (0 = \tau_0 < \tau_1 < ... < \tau_{K-1} < \tau_K = 1), \ \ \tau_k - \tau_{k-1} \geq \Delta_\tau, \forall k = 1, ..., K\}\,. \tag{4.12}$$

The penalised maximum-likelihood estimate may be written

$$(\hat{K}, \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\lambda}}) = \arg \min_{K=1,...,K_{\max}} \min_{\boldsymbol{\tau}\in\mathcal{T}^\Delta_K, \boldsymbol{\lambda}\in\Theta^K} \tilde{\mathcal{C}}_{Y_{1:n}}(K, \lfloor n\boldsymbol{\tau}\rfloor, \boldsymbol{\lambda}) \ , \tag{4.13}$$

with $\tilde{\mathcal{C}}_{Y_{1:n}}(K, \lfloor n\boldsymbol{\tau}\rfloor, \boldsymbol{\lambda})$ defined as in Eq. 4.11.

Consistency of the penalised estimate may be shown by introducing one more hypothesis.

(H5) The sequence β_n verifies $\beta_n \xrightarrow[n\to\infty]{} \infty$ and $\frac{\beta_n}{n} \xrightarrow[n\to\infty]{} 0$.

The latter hypothesis is checked by a large palette of penalties, including the BIC criterion, $\beta_n = \frac{\ln n}{2}$.

Proposition 4.1.3. *Under the hypotheses (H1)-(H5), the penalised maximum-likelihood estimate defined in Eq. 4.13 is weakly consistent*

$$(\hat{K}, \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\lambda}}) \xrightarrow[n\to\infty]{\mathbb{P}^\star} (K^\star, \boldsymbol{\tau}^\star, \boldsymbol{\lambda}^\star) \ . \tag{4.14}$$

The proof of this assertion is a direct consequence of Theorem 3.1 in [53].

4.2. Estimation procedure

The model having been introduced and its theoretical properties established, one may now focus on its practical implementation. As the true number of change-points is seldom known for real datasets, the proposed algorithm considers the case where $K^\star$ is unknown.

4.2.1. Estimating the λ's

First of all, one should be able to properly compute $\lambda_{s:t}$ as the estimate of the left-censored exponential parameter, given the sub-sample $Y_s, \ldots, Y_t$. After having written the log-likelihood similarly to Eq. 4.7, and computed the derivative with respect to λ, one may notice that the latter does not lead to an analytical expression for $\lambda_{s:t}$:

$$\frac{\partial \sum_{i=t_k+1}^{t_{k+1}} \ln f(y;\lambda,\sigma)}{\partial \lambda} = \sum_{i=t_k+1}^{t_{k+1}} \frac{a^\sigma \sigma(\lambda)^{\sigma-1} e^{-(\lambda a)^\sigma}}{1-e^{-(\lambda a)^\sigma}} \mathbb{1}_{\{y=a\}} + \sum_{i=t_k+1}^{t_{k+1}} \left(\frac{\sigma}{\lambda} - \sigma(\lambda)^{\sigma-1} y^\sigma\right) \mathbb{1}_{\{y>a\}}, \tag{4.15}$$

Hence, instead of having an explicit formula for $\lambda_{s:t}$, one derives it through a numerical optimisation procedure. In the work presented here, the Newton-Raphson method was used to search for the zeros of the first derivate of the cost function in order to find its minimum. Furthermore, checking that the second derivative is strictly positive, thus guaranteeing the unicity of the maximum likelihood estimate, can prove to be a difficult task. This study is made in Appendix A.
However, the case where all data in the sample $Y_s, \ldots, Y_t$ are censored is an exception. Looking at the analytical likelihood formula, we find that the λ realising the minimum of the cost function still exists, but tends toward infinity. This result is shown in Figure 4.1 where We can see it is the case for any σ value. To solve this problem from a computational point of view, we impose an upper bound on the possible values of lambda. In the remainder of this work, the maximum value of λ is set to 10^6. When the value of the estimate exceeds this threshold, the Newton-Raphson procedure terminates.

4.2.2. Estimation procedure for $K^\star$, $\boldsymbol{t}^\star$ and $\boldsymbol{\lambda}^\star$

The estimation procedure aims at minimising the penalised log-likelihood in Eq. 4.11. The optimisation problem is equivalent to finding both the best partitioning of the data as defined

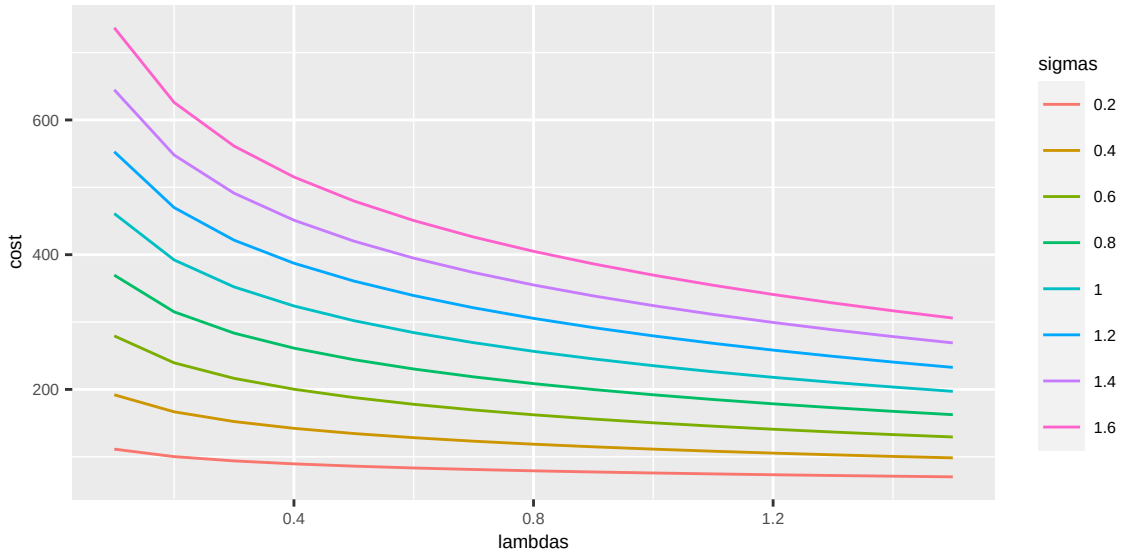

Figure 4.1: Plot of the cost function values against λ values when all observations are censored. It is represented for several σ values. The sample consists in 100 values of threshold $a = 0.1$.

by $\hat{K}$ and $\hat{\boldsymbol{t}}$, as well as the estimates for the scale parameters of the left-censored Weibull within each segment, $\hat{\boldsymbol{\lambda}}$. the complexity related to the search of the optimal partitioning is prohibitive in practice, several heuristics based on dynamic programming have been proposed in the literature [65, 47, 48, 93]. The approach adopted here was the Pruned Exact Linear Time algorithm [52] as introduced in Chapter 2, which has the advantage of achieving a linear complexity in the number of data. Let's denote $F(s)$ for $s \in [1:n]$ the best segmentation found so far on the data $(y_1, ..., y_s)$:

$$F(s) = \max_{\boldsymbol{\tau} \in \mathcal{T}_s} \left(\sum_{k=1}^{m+1} [W(Y_{(t_k+1):t_{k+1}}, \lambda_k) - \beta_n] \right),$$

with $\mathcal{T}_s = \{\boldsymbol{t} : 0 = t_0 < t_1 < ... < t_m < t_{m+1} = s\}$ be the set of all possible segmentation of signal $y_{1:s}$. With these notations, the procedure may be written as illustrated in Algorithm 4. The resulting algorithm is very similar to the one exposed in Chapter 2. In Algorithm 4 however, the $\hat{\lambda}_{(t+1):\tilde{t}}$ is the output of the Newton-Raphson method, trained on $y_{t+1}, ..., y_{\tilde{t}}$. Let us mention here that Algorithm 4 relies on a series of hyper-parameters which ought to be tuned: the precision threshold and the maximum number of iterations in the Newton-Raphson step, the initial values for the λ's in Newton-Raphson step also, and the minimum number of observations between two consecutive change-points.

4.3. Simulation study

The model presented in the previous sections is subjected to a series of tests. The purpose of these tests is twofold:

Algorithm 4 PELT algorithm combined with Newton-Raphson:

input : the data $y_1, ..., y_n$, the censoring threshold a, and the penalty term β_n

initialisations : $F(0) = \beta_n$, $R_1 = \{0\}$, $CP(0) = NULL$

for all $\tilde{t} = 1, ..., n$ **do** :

1. Compute $F(\tilde{t}) = \min_{t \in R_{\tilde{t}}} \{F(t) + W(y_{(t+1):\tilde{t}}, \hat{\lambda}_{(t+1):\tilde{t}}) + \beta_n\}$
2. Compute $\bar{t} = \arg\min_{t \in R_{\tilde{t}}} \{F(t) + W(y_{(t+1):\tilde{t}}, \hat{\lambda}_{(t+1):\tilde{t}}) + \beta_n\}$
3. Set $CP(\tilde{t}) = [CP(\bar{t}), \bar{t}]$
4. Set $R_{\tilde{t}+1} = \left\{ t \in R_{\tilde{t}} \cup \{\tilde{t}\} | F(t) + W(y_{(t+1):\tilde{t}}, \hat{\lambda}_{(t+1):\tilde{t}}) + \beta_n \leq F(\tilde{t}) \right\}$

end for

output : the vector of change-points CP.

1. As mentioned in Section 4.2, our algorithm is driven by hyperparameters. We would like to be able to adjust these to achieve good performance.
2. We would also like to develop a procedure to calibrate the penalty value and compare its performance with a method from the litterature.

To assess the good tuning and efficiency of our method, we will compare its performances with the *MultRank* method developped by [63] presented in Chapter 2. Since this method being also adapted for censored data, it constitutes a coherent reference point.

4.3.1. Tuning hyper-parameters of the Newton-Raphson method

- **Precision of the Newton-Raphson method:** the precision criterion in the Newton-Raphson algorithm implemented here is user-defined threshold ϵ. If we denote $\hat{\lambda}_n$ the estimate obtained at the n-th iteration in the Newton-Raphson loop, the criterion can be written $|\hat{\lambda}_{n+1} - \hat{\lambda}_n| \leq \epsilon$. In other word, the method stops when the correction made to the estimate value at the n-th step is below ϵ. The threshold ϵ is fixed at the value 10^{-8}. (CITER UNe SOURCE POUR LES NORMES IEEE ?)
- **Initialisation value in the Newton-Raphson method:** four initialisation values are available in our implementation. We can choose between the classical techniques such as the moment method estimator λ_{init}^{MM} [49], the quantile inversion estimator λ_{init}^{QI}, the weighed maximum likelihood estimator λ_{init}^{WMLE} [84] or the classical maximum likelihood estimator λ_{init}^{MLE} of a Weibull scale parameter. Supposing a sample of observations $\boldsymbol{x} = (x_1, ..., x_n)$ generated from a left censored Weibull of parameters (λ, σ) and censoring threshold a, we can define them as follow :
  - $\lambda_{init}^{MM} = \frac{\Gamma(1+\frac{1}{\sigma})}{\bar{\boldsymbol{x}}}$

- $\lambda_{init}^{QI} = \frac{\left(-\ln(1-\alpha)\right)^{\frac{1}{\sigma}}}{q_{\boldsymbol{x}}^{\alpha}}$
- $\lambda_{init}^{WMLE} = \left(\frac{1}{n q_{\mathcal{W}(n,n)}^{0.5}} \sum_{i=1}^{n} x_i^{\sigma}\right)^{-\frac{1}{\sigma}}$
- $\lambda_{init}^{MLE} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\sigma}\right)^{-\frac{1}{\sigma}}$,

where $q_{\mathcal{W}(n,n)}^{0.5}$ is the median of Weibull with parameters (n, n), $q_{\boldsymbol{x}}^{\alpha}$ is the α-th empirical quantile of the sample $\boldsymbol{x}$.
Two important points must be noted. First, all the initialisation values depend on σ. It is not problematic in our simulation tests because it is supposed known and fixed. However it stresses again the necessity of its estimation in the future (see Chapter 5). Second, those estimators do not take the censorship into account. They are all biased (except if the sample $\boldsymbol{x}$ does not bear any censored values).
We tested all possible configurations with the varying values of $n = (20, 100, 500)$, $\lambda = (1/100, 1, 100)$ and a depending on a censoring rate $\alpha = (0.05, 0.25, 0.5, 0.75, 0.95)$. a was the threshold such that $\alpha\%$ of the sample was censored. The shape parameter is supposed known and fixed at $\sigma = 0.5$. For each cases, we simulated $N = 1000$ samples of left censored Weibull with shape parameter λ and censored rate α. We then compute the mean of all estimates for each initialisation values. All the results are stored into Tables 4.1 and 4.2. The simulations show that all initialisation values lead to extremely similar results. It is worth mentioning that the quantile is not reliable for low values of n. In the rest of this work, the initialisation value will be defined as the weighed maximum likelihood. The table for the case where $n = 500$ is not displayed because all methods gave the same results. However, the most important result of this experiment is that the method converges and that the choice of the initialization point is important. From now on, we choose to initialize the method with the weighed Maximum Likelihood Estimator. More experiences are provided in Appendix A.

■ **Maximum number of iteration in the Newton-Raphson method N_{max}:** this parameter is user-defined. When the Newton-Raphson method reaches N_{max}-th iteration, it stops. The following experience has been conducted to tune this parameter :

1. Generate a n sized sample $y_1, ..., y_n$ following a left-censored Weibull with scale and shape parameters λ and σ and with censoring threshold a censoring $\alpha\%$ of the sample. We chose $n = 100$.
2. Store the minimum iteration such that the estimate value $\hat{\lambda}$ does not evolve. The maximum iteration value allowed for the experiment is $N_{max} = 100$
3. Repeat the two first step $N = 1000$ times.
4. Compute the mean value of the minimum iteration and the mean values of the N samples estimated parameters.

Several configurations are tested with different values of threshold α and λ. The shape parameter is fixed at $\sigma = 0.5$. We initialize the method with the Weighed Maximum Likelihood value. The results are stored in Table 4.3. Looking at the average minimum iteration needed to reach a stable value, N_{max} was set to 100 which seems a reasonable choice for the rest of this work. It can be noted that the minimum number of iterations needed for low values of λ are higher than in other configurations

Summary of the simulations : we calibrate the Newton-Raphson method the following way:

- Precision threshold $\epsilon = 10^{-8}$.
- Maximum number of iterations tolerated $N_{max} = 100$.
- Initialisation method : weighed Maximum Likelihood Estimator.

α	λ	$\overline{\hat{\lambda}}_{WMLE}$	$\overline{\hat{\lambda}}_{MLE}$	$\overline{\hat{\lambda}}_{QI}$	$\overline{\hat{\lambda}}_{MM}$
0.05	100.00	116.77	116.77	1963.08	116.77
0.05	1.00	1.16	1.16	1.47	1.16
0.05	0.01	0.01	0.01	1790.22	0.01
0.25	100.00	119.24	119.24	151.37	119.24
0.25	1.00	1.16	1.16	2362.08	1.16
0.25	0.01	0.01	0.01	6038.39	0.01
0.50	100.00	118.07	118.07	3687.48	118.07
0.50	1.00	1.18	1.18	1.48	1.18
0.50	0.01	0.01	0.01	0.02	0.01
0.75	100.00	122.44	122.44	1724.86	122.44
0.75	1.00	1.25	1.25	2602.37	1.25
0.75	0.01	0.01	0.01	1189.03	0.01
0.95	100.00	163.51	163.51	165.04	163.51
0.95	1.00	1.62	1.62	1.63	1.62
0.95	0.01	0.02	0.02	0.02	0.02

Table 4.1: Choice of initialisitation value: simulation results for $n = 20$.

4.3.2. Tuning the parameters of the change-point detection method

- **The minimal segment length:** this argument is implicitly introduced when the number of changes $K^\star$ is not known. We want to calibrate this parameter by comparing the ability of our method to detect the presence of a change point in the data with the *Multrank* method. In this context, comparing the non-parametric approach with the parametric approach is equivalent to using a likelihood ratio for the latter. It should be noted that performing a likelihood ratio test or maximising the penalised likelihood introduced in the Equation 4.13 for $K_{max} = 1$ is equivalent whatever the choice of penalty might be. The

α	λ	$\overline{\hat{\lambda}}_{WMLE}$	$\overline{\hat{\lambda}}_{MLE}$	$\overline{\hat{\lambda}}_{QI}$	$\overline{\hat{\lambda}}_{MM}$
0.05	100.00	102.65	102.65	102.98	102.65
0.05	1.00	1.03	1.03	1.03	1.03
0.05	0.01	0.01	0.01	0.01	0.01
0.25	100.00	102.97	102.97	103.52	102.97
0.25	1.00	1.03	1.03	1.03	1.03
0.25	0.01	0.01	0.01	0.01	0.01
0.50	100.00	104.23	104.23	104.40	104.23
0.50	1.00	1.03	1.03	1.03	1.03
0.50	0.01	0.01	0.01	0.01	0.01
0.75	100.00	104.41	104.41	104.49	104.41
0.75	1.00	1.04	1.04	1.04	1.04
0.75	0.01	0.01	0.01	0.01	0.01
0.95	100.00	110.90	110.90	110.90	110.90
0.95	1.00	1.10	1.10	1.10	1.10
0.95	0.01	0.01	0.01	0.01	0.01

Table 4.2: Choice of initialisitation value: simulation results for $n = 100$.

	$n = 20$			$n = 100$			$n = 500$		
α \ λ	100	1	0.01	100	1	0.01	100	1	0.01
0.05	9.15	9.38	12.31	8.61	8.75	25.45	8.24	8.19	9.65
0.25	9.57	9.60	48.05	8.44	9.33	15.38	7.98	9.15	8.87
0.50	10.35	10.82	43.56	10.28	10.57	53.29	10.55	10.09	33.27
0.75	11.18	11.77	27.64	11.19	11.51	12.20	10.90	11.84	11.28
0.95	13.65	13.91	26.93	14.01	14.01	39.60	14.15	14.18	45.00

Table 4.3: Choice of maximum number of iteration N_{max}: simulation results.

statistics of the non-parametric test and the likelihood ratio test are calculated for each sample and will allow the calculation of the ROC curves [27] and the corresponding areas under the curve (AUC) to compare the performances of the two approaches. The results obtained on the simulated data are shown in Figure 4.2. For both methods, the area under the ROC curve is calculated by varying the minimum number of observations before a change point. According to these results, the parametric method performs better when the interval between two interruptions contains enough observations. In addition, the performance of the two methods is also compared as a function of the censoring threshold. We choose to fix the minimum segment length value to 50 observations. We are assured that in a highly censored context the parametric method will perform correctly.

■ **The penalty calibration:** The last parameter to be set is the penalty β. While calibrating β, we want to compare the performance of the break detection with the Multrank method. The experimental framework is as follows:

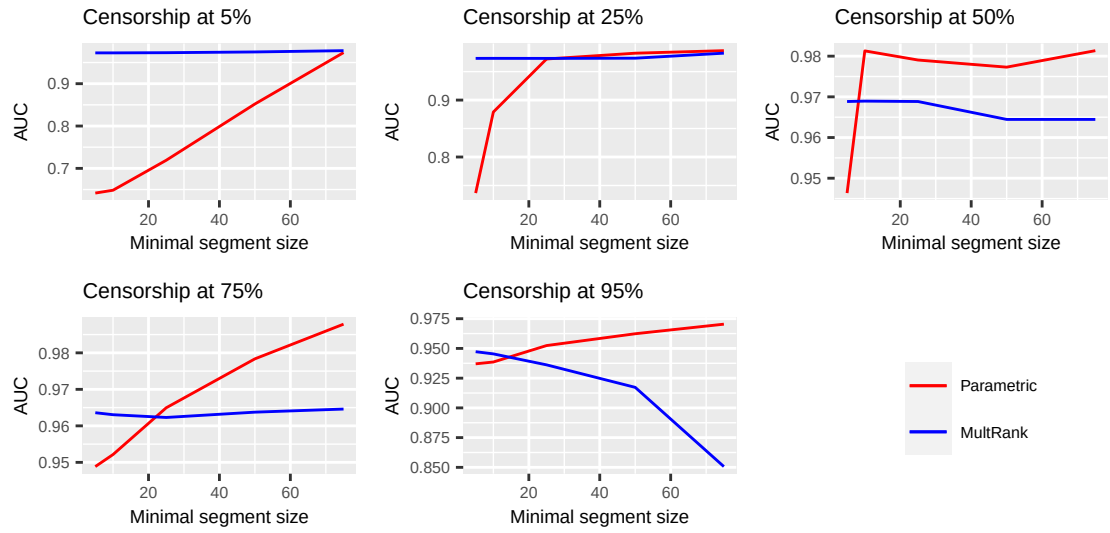

Figure 4.2: Choice of the minimal segment length: simulation results. Our method performance is Illustrated with the red line, the *Multrank* method is drawn in blue. The results are illustrated for several censorship thresholds and the different minimal segment lengths used were $5, 10, 25, 50$ and 75 observations.

1. we simulate $N = 100$ samples $(x_1, ..., x_n)$ of size $n = 400$ following a left-censored Weibull distribution with $\alpha\%$ of censored data. We made tests for the different censorship rates $\alpha = (25, 50, 75, 95)$. The shape parameter of the Weibull distribution is assumed to be known and set to $\sigma = 0.5$. The scaling parameters $\boldsymbol{\lambda}^\star$ have $K^\star = 4$ breaks at positions $p_1^\star = 80$, $p_2^\star = 160$, $p_3^\star = 240$ and $p_4^\star = 320$ and take the values $\boldsymbol{\lambda}^\star = (\lambda_1^\star = 1, \lambda_2^\star = 4, \lambda_3^\star = 0.5, \lambda_4^\star = 5, \lambda_5^\star = 1)$. An example of a sample simulated in this way is shown in Figure 4.3.
2. For each of the N samples, we perform the parametric change-point detection and the Multrank methods. For each sample, we obtain the estimated number of breaks $\hat{K}_{param}$ and $\hat{K}_{multrank}$ and their position $(\hat{p}_{k,param})_{k=1}^{\hat{K}_{param}}$ (respectively $(\hat{p}_{k,multrank})_{k=1}^{\hat{K}_{multrank}}$).
3. for both methods, we count the number of samples among the N for which the correct number of breaks has been estimated (e.g. $\hat{K}_{param} = K^\star$). Also, for each of the samples for which the estimate of $K^\star$ is correct, we examine the distance between the estimated position of a change-point and its nearest true break $\min_{k,i\in[1:K^*]}(\hat{p}_k - p_i^\star)$.

In the case where K is not known, we proceed as follows for each method to estimate it:

– For the parametric method: we use the algorithm CROPS, algorithm to scan a continuous range of penalty values $[\beta_{min}, \beta_{max}]$. We obtain a set of B values $(\hat{\beta}_1, ..., \beta_B)$ and the optimal segmentations associated with these penalty values. We then plot the cost of the segmentations as a function of the number of breaks. We choose the optimal penalty using a elbow heuristic. This procedure is described in [38]. The choice of β_{min} and β_{max} is inspired from linear penalties like the BIC criterion [100].

Note that when using the BIC penalty in change point detection, the penalty term written in section 4.2.2 becomes : $\beta_n = \frac{D}{2}\log(n) = \frac{1}{2}\log(n)$, where D is the number of dimensions of the parameter. More precisely, we took a wide interval of penalty values defined by $\beta_{min} = \frac{\log(n)}{10}$ and $\beta_{max} = 5\log(n)$.

– For the non parametric *Multrank* method, we compute the optimal segmentation for k breaks, where k ranges from 1 to K_{max}. For each of these segmentations, we can compute the value of the statistic $I_k(n)$ statistic described in [64]. As in the parametric method, we represent the values of this statistic as a function of k, and we determine the number of estimated breaks by an elbow heuristic. Here, K_{max} is fixed at $2 * K^\star = 8$.

The results of the simulations are shown in Table 4.4 and in Figure 4.4. It can be seen that in the ideal scenario, where the data are indeed distributed according to a left-censored Weibull distribution, the parametric method performs better both in detecting the correct number of breaks and in accurately estimating their position. However, this performance decreases as the censoring rate increases.

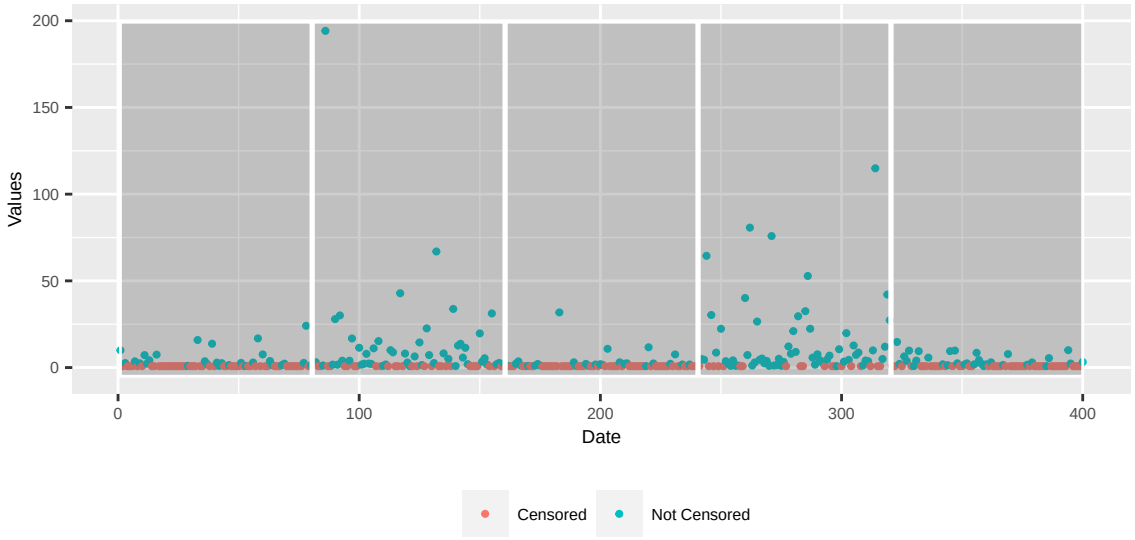

Figure 4.3: Example of simulated signal with $(\lambda_1 = 1, \lambda_2 = 4, \lambda_3 = 0.5, \lambda_4 = 5, \lambda_5 = 1)$, $\sigma = 0.5$, $n = 400$, $K = 4$, $(p_1 = 80, p_2 = 160, p_3 = 240, p_4 = 320)$ and $\alpha = 50\%$.

$\alpha(\%)$	Parametric method	MultRank
25	84	58
50	80	63
75	87	68
95	65	10

Table 4.4: Number of correct estimations of K over $N = 100$ samples for both methods for different $\alpha\%$ censorship rates.

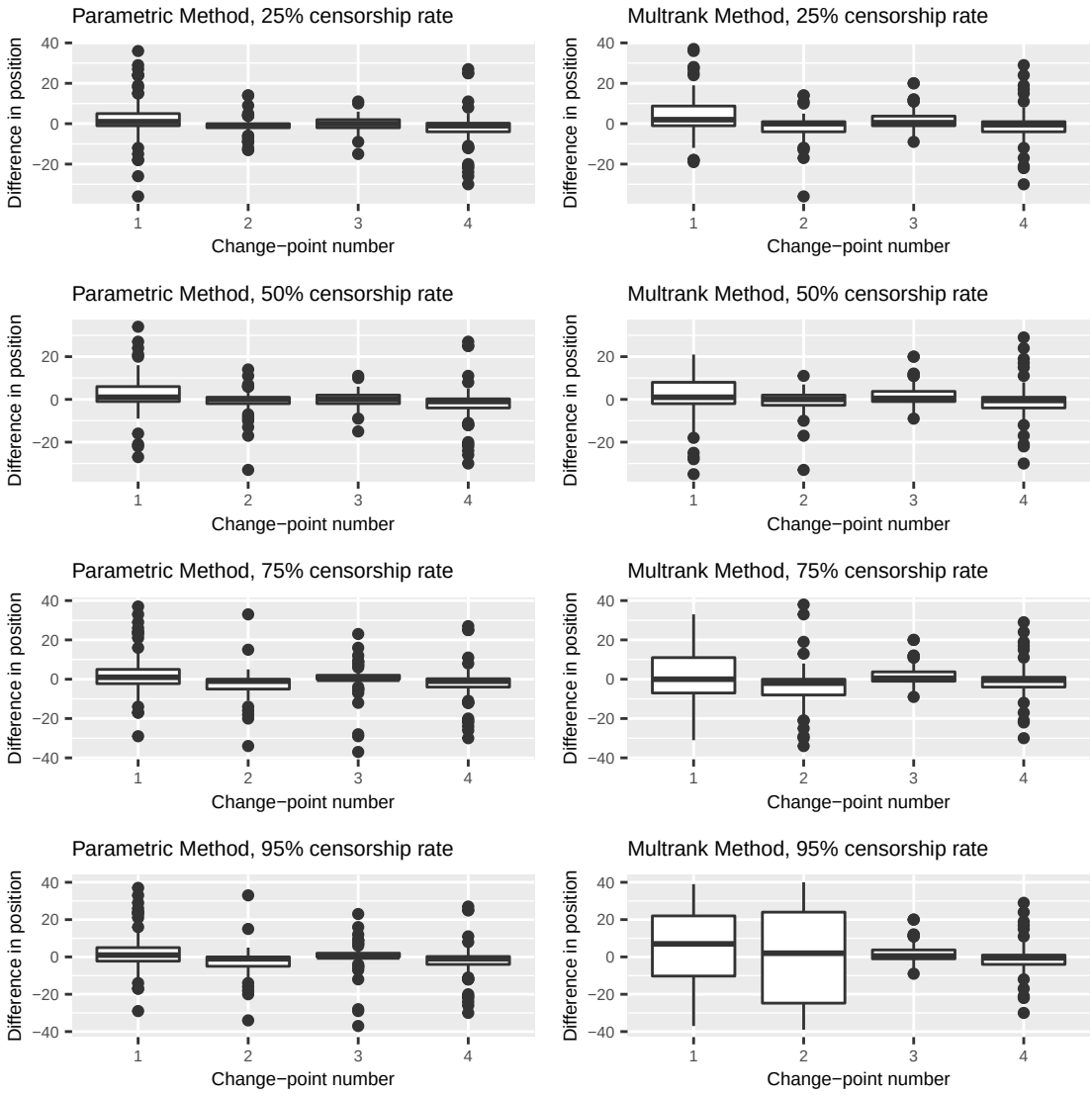

Figure 4.4: Precision of the estimated change-points for both methods.

5. Chapter 5 : Case study

Contents

5.1 Data collection procedure and associated generative model 48
5.1.1 Monitoring stations network 48
5.1.2 Data collection 49
5.1.3 A piece-wise stationary model for the coarse-grain time series 50
5.2 Methods 50
5.2.1 Piece-wise stationary model estimation via change-point detection . . 50
5.2.2 Spatial clustering 52
5.2.3 Anomaly detection 53
5.3 Data presentation 55
5.3.1 Time period and geographical area selection 55
5.3.2 Graphical representation of the station network 56
5.4 Results 57
5.4.1 Temporal segmentation 57
5.4.2 Spatial segmentation 59
5.4.3 Anomalous cluster identification 61

Monitoring the environmental pollution is of great interest for public authorities, important adverse health-effects being well documented nowadays [50, 68, 73]. National health agencies are thus much concerned with monitoring ambient levels and quantifying the concentration of various pollutants in given environmental areas.
At the same time, modelling environmental pollution data is a complex issue, due to several reasons, some intrinsic to the types of data under study, some specific to the data collection process implemented in different countries. Firstly, pollutant concentration levels are measured by sensors which have generally detection and quantification limits: the corresponding data are then left-censored. Secondly, the data is usually skewed to the right, with long tails hinting high concentrations. Thirdly, in numerous situations the data is irregularly sampled because of measurement practices, and is often multivariate, since various pollutant levels are monitored. Fourthly, pollution is monitored in various locations, each location possibly using different sensors, yielding a significant spatial heterogeneity. Notice that the last two problems are specific to some applications such as the one considered in the present paper. Indeed many countries have implemented very strict data collection protocols that ensure regular measurement rates on standardized sensors for a selection of pollutants[1].

[1]See for example the air quality data reported by UK-AIR `https://uk-air.defra.gov.uk/`.

How to handle left-censored and right-skewed data is therefore one of the first aspects to consider when modelling environmental data. A rich literature has been developed on this topic during the last thirty years, and may be roughly divided into three categories of approaches: substitution methods (censored data is imputed using some values chosen *a priori* or via a generative model), parametric methods (maximum likelihood estimates are computed under the hypothesis that the data comes from some log-normal, Weibull, Gamma, exponential, or other log-logistic distribution), and non-parametric methods (Kaplan-Meier or hazard function estimates). Detailed reviews of the various approaches are available for instance in [26, 42, 69, 14, 4, 34, 89].

The second aspect to consider is spatio-temporal heterogeneity. Air pollution data has received, for instance, a great deal of attention, and several modelling approaches have been proposed in the literature. Some are based on temporal regression models combined with kriging [85, 60], while others use latent variables and co-clustering approaches [11]. Nevertheless, these approaches do not include the fact that monitoring data is not normally distributed, and is usually left-censored. In the specific field of pesticide concentration monitoring, several recent papers address the spatio-temporal issue from an exploratory point of view see for instance [15, 29, 6].

If one focuses specifically on temporal heterogeneity, a common approach to deal with it is to use a change-point based segmentation. Assuming the data is strictly stationary, conditionally to a (possibly unknown) number of change-points and their associated locations, change-point analysis aims at identifying the number of change-points (also known as *breaks*), their locations, and the characteristics of the probability distribution within each temporal segment. Widely used in a variety of applications [9, 16, 62, 80, 54], and, in particular, for environmental pollution monitoring [20], change-point detection is a reference technique for time series segmentation. The present chapter relates to the offline framework, by supposing the full data has been recorded, and the segmentation is done posterior wise. For recent and detailed reviews of the offline change-point detection, the reader may refer to [92, 8]. While the literature on change-point detection is abundant, applications to spatial data are somewhat limited. An early example of such method can be found in [66] while recent advances in a setting close to ours are presented in [17]. As far as we know, none of the existing change-point detection method for spatial data applies to irregularly sampled and sparse data (on the temporal axis). In this chapter, we tackle the issue of pesticide concentration monitoring, and introduces a new methodology which integrates both the specific left-censored distribution of the data, and the spatio-temporal context. The main goal is to identify contextual anomalies, both from a temporal and a spatial point of view. The proposed method builds on a parametric model for left-censored and right-skewed distributions, and combines it with a change-point detection step and a clustering step.

Change-point detection is used for modelling temporal heterogeneity, by assuming a piece-wise stationary distribution on the series of maximum values, for a given time resolution. It produces temporal segments in which the pesticide concentrations are assumed to follow a stationary distribution.

Clustering is then used for modelling the expected spatial homogeneity while integrating geographical constraints such as river networks, wind directions, etc. Indeed, as geological, terrain and climatic characteristics of an area can influence the dispersion of a chemical substance and

on its potential use in the case of e.g. a pesticide, concentrations are expected to be somewhat correlated in small scale regions that are homogeneous in terms of influencing characteristics. Especially in the application presented here, which relates to the investigation of pollutants in surface waters, it is interesting to take into account the hydrographic structure of the region as in e.g. [17]. Indeed, if a high concentration of a substance is detected at a certain point in time, traces of this substance should be found later downstream. This hypothesis is accounted for by building clusters of measuring stations according to their proximities measured via the hydrographic network.
Conditionally to the temporal segment detected by the change-point procedure, and to the spatial cluster detected by the clustering procedure, one may analyse the data and identify contextual anomalies.
The rest of the chapter is organised as follows: in Section 5.1, the generative model assumed for environmental pesticide monitoring data is described; the proposed method for estimating and handling this model from observed data is detailed in Section 5.2; a detailed example on data collected by French authorities in Val de Loire region is fully illustrated in Sections 5.3 and 5.4.

5.1. Data collection procedure and associated generative model

We study specifically in this chapter a non homogeneous data collection process for pesticide use monitoring. It is represented by a generative model with two levels. The first level, a.k.a. the fine-grain level, consists of a network made of monitoring stations, where each station is associated to an irregularly sampled time-series. The second level, a.k.a. the coarse-grain level, summarises the maximum recorded values throughout the network, for a specified temporal resolution, and assumes a piece-wise stationary distribution.

5.1.1. Monitoring stations network

We consider a network of monitoring stations used to collect concentration measurements at irregularly sampled instants. The stations are represented by an undirected graph $G = (V, E)$, which vertices $V = (v_i)_{1 \leq i \leq N}$ are the monitoring stations and which weighted edges E are links between stations that are directly comparable. The aim of the graph is to represent expert knowledge about expected measurement homogeneity. When two stations are connected in G, their measurements can be compared directly: a small edge weight assumes simultaneous measurements to be close, while a large one allows for significant differences. Shortest paths in the graph can be used to compare stations that are not directly connected, using the total weight of the paths to measure non homogeneity. This approach is inspired by methods developed for signal processing on graphs [90], but we use a dissimilarity based weighting rather than the classical similarity based one.
This graph based representation is very flexible and can be used to model different types of spatial homogeneity. For instance, the focus of the present paper is the monitoring of water

concentration of pesticides and thus dissimilarities between stations will be computed based on the network of rivers on which they are situated (see Section 5.2.2). Other modelling approaches may use a different graph considering for instance dominant wind directions relevant for air diffusion of pollutants.
This graph is not necessarily fully connected, there can be P non connected components that we will denote $(\mathcal{K}_1, ..., \mathcal{K}_P)$.

5.1.2. Data collection

Each station v_i is supposed to be associated to a time series $(y_{ij}, t_{ij})_{1 \leq j \leq p_i}$, where p_i is the number of sampled data points at v_i, and y_{ij} is the concentration level of some pollutant at time t_{ij}. All measurements y_{ij} are left-censored by some threshold q_{ij}, representing the quantification limit. Quantification limits depend on the machines used at each station and at each time instant, hence depend both on the station v_i and on the collection instant t_{ij}. Furthermore, quantification limits are supposed to be known, fixed quantities.
Summarising the above notations and hypotheses, a data set sampled from the stations network is given by a collection of measurements and associated quantification limits, and denoted

$$\mathcal{D} = \Big((y_{ij}, t_{ij}, q_{ij})_{1 \leq j \leq p_i} \Big)_{1 \leq i \leq N} .$$

Notice that in practical applications, we expect to have a rather small number of measurements for each station, i.e. to have small values for the p_i. In addition, we do not expect the measurement instants to be shared among the stations. See Section 5.3.1 for examples.
From the complete representation of the data $\mathcal{D}$, one may derive an aggregated, coarser representation. First, an adapted temporal resolution for the phenomenon at study is selected. For instance, in the case of the present study, a daily resolution is considered. Second, the selected resolution is used to build a time series of increasing instants $(\tau_k)_{1 \leq k \leq K}$, at which at least one observation is available in the data collection. We denote $t_{ij} \in \tau_k$ the fact that the observation time t_{ij} is compatible with τ_k at the specified resolution, e.g. that the observation y_{ij} was made during the day τ_k.
Third, once $(\tau_k)_{1 \leq k \leq K}$ has been computed, one may introduce a coarse-grain, global series, summarising the maximum values recorded within the temporal resolution with

$$\overline{y}_k = \max \{ y_{ij} \mid t_{ij} \in \tau_k \} . \tag{5.1}$$

For instance, for a daily aggregation level, $\overline{y}_k$ is the largest value among all the measurements that took place during day τ_k. Notice that $(\overline{y}_k)_{1 \leq k \leq K}$ is left-censored as the consequence of the censoring of the underlying values. The quantification limit for $\overline{y}_k$ is denoted $\overline{q}_k$, with

$$\overline{q}_k = \max \{ q_{ij} \mid t_{ij} \in \tau_k \} . \tag{5.2}$$

The coarse representation of $\mathcal{D}$ is then

$$\overline{\mathcal{D}} = (\overline{y}_k, \tau_k, \overline{q}_k)_{1 \leq k \leq K} . \tag{5.3}$$

5.1.3. A piece-wise stationary model for the coarse-grain time series

In order to model the global use of the substance under monitoring, a piece-wise stationary generative model is introduced for the coarse data set $\overline{\mathcal{D}}$. The model is based on the following assumptions:

- there are $L^* > 0$ change-points producing $L^* + 1$ stationary intervals defined by

$$0 = \eta_0^* < \eta_1^* < \ldots < \eta_{L^*}^* < \eta_{L^*+1}^* = K;$$

- the observations $(\overline{y}_k)_{1 \leq k \leq K}$ are realisations of K independent random variables $(\overline{Y}_k)_{1 \leq k \leq K}$;
- when $k \in [\eta_{l-1}^* + 1, \eta_l^*]$, $\overline{Y}_k$ is distributed according to a left-censored parametric Weibull distribution with interval dependent parameters λ_l^* and a left-censoring threshold $\overline{q}_k$, which is a known constant. The shape parameter σ^* is supposed unknown and fixed throughout the whole signal.

Several remarks must be pointed out at this point. First, notice that the model only accounts for the concentrations $\overline{y}_k$ but not for the instants and the quantification limits which are supposed deterministic quantities. The second remark is that the $(\eta_l)_{l=1}^{L^*+1}$ define the **contextual** aspect for the anomaly detection step implemented in 5.2.3.

5.2. Methods

We are interested in finding anomalies in data collected according to this spatiotemporal model. The proposed methodology combines two different homogeneity models. The temporal aspect is based on the piece-wise stationary model proposed in Section 5.1.3, while the spatial aspect is based on the graphical representation introduced in Section 5.1.1. In a first step, we estimate the parameters of the temporal model. In a second step, the homogeneity assumptions represented by the graph of stations is used to detect stations with anomalous measurements with respect to close stations in a given stationary temporal segment.

5.2.1. Piece-wise stationary model estimation via change-point detection

The coarse-grained data $\overline{\mathcal{D}}$ is segmented using a change-point detection approach applied to the model introduced in Section 5.1.3. Since both the number and the location of the change-points are unknown, we shall optimise a penalised cost function, and seek to estimate the number of change-points L^*, the change-point locations $\boldsymbol{\eta}^\star = (\eta_l^*)_{1 \leq l \leq L^*}$, the parameters $\boldsymbol{\lambda}^\star = (\lambda_l^*)_{1 \leq l \leq L^*+1}$ and σ^*.
The estimates write as

$$(\widehat{L}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\lambda}}, \widehat{\sigma}) = \arg\min_{L, \boldsymbol{\eta}, \boldsymbol{\lambda}} \mathcal{C}(L, \boldsymbol{\eta}, \boldsymbol{\lambda}, \sigma; \overline{\mathcal{D}}). \tag{5.4}$$

The penalised cost is given by

$$\mathcal{C}(L, \boldsymbol{\eta}, \boldsymbol{\lambda}, \sigma; \overline{\mathcal{D}}) = \sum_{l=1}^{L+1} -\ln \mathcal{L}(\lambda_l, \sigma; \overline{y}_{\eta_{l-1}+1}, \ldots, \overline{y}_{\eta_l}) + \beta_K (L+1) D, \tag{5.5}$$

where $\mathcal{L}(\lambda_l, \sigma; \overline{y}_{\eta_{l-1}+1}, \ldots, \overline{y}_{\eta_l})$ is the likelihood of the l-th segment for the left censored Weibull distribution with parameters (λ_l^*, σ^*), β_K is the penalty to apply at the addition of new segment, and D is the dimension of the parameter vectors λ_l.
For fixed values of σ, L and of $\boldsymbol{\eta}$, $\mathcal{C}(L, \boldsymbol{\eta}, \boldsymbol{\lambda}, \sigma; \overline{\mathcal{D}})$ is maximized by setting $\boldsymbol{\lambda}$ to the maximum likelihood estimate (MLE), $\widehat{\boldsymbol{\lambda}}_{MLE}(L, \boldsymbol{\lambda})$. Thus, for a fixed value of σ, the optimisation problem may be further written as

$$(\widehat{L}, \widehat{\boldsymbol{\eta}}) = \arg\min_{L, \boldsymbol{\eta}} \mathcal{C}(L, \boldsymbol{\eta}, \widehat{\boldsymbol{\lambda}}_{MLE}(L, \boldsymbol{\eta}), \sigma; \overline{\mathcal{D}}). \tag{5.6}$$

We can see that this procedure is close to the one proposed in Chapter 4. The main difference is that σ is not known. In order to limit the number of parameters to estimate, and also to avoid numerical issues rapidly induced by the large number of censored data, the shape parameter σ in the Weibull distributions is supposed not to vary with the change-points. σ is thus constant throughout the series, and is estimated globally under a stationary hypothesis. The only parameter supposed to be varying at each change-point is therefore the rate of the Weibull distribution, say λ. Furthermore, we add the assumption that $\sigma \leq 1$. This comes from the observation of the signal $\overline{\mathcal{D}}$ presented in 5.3. It serves the purpose of modelling heavy tailed distribution.
We propose the following heuristic to estimate σ and to calibrate the penalty term β_n. It consists in an iterative procedure where we alternate between optimizing 5.6 with respect to σ and (L, η). The goal is to construct a sequence $\widehat{\boldsymbol{\sigma}} = (\widehat{\sigma})_{r=0}^{R}$ that converges to σ^* when $R \to +\infty$ and then to retrieve the results of the best segmentation possible for the last $\widehat{\sigma}_r$ computed. In the initialisation step, the estimate of σ is computed from the whole signal $(\overline{y}_1, ..., \overline{y}_K)$ by supposing that there is no change-point. More formally, it is equivalent to take the case where $L = 0$ and minimizing the cost:

$$\mathcal{C}(\lambda, \sigma; \overline{\mathcal{D}}) = -\ln \mathcal{L}(\lambda, \sigma; \overline{y}_1, \ldots, \overline{y}_K) + \beta_K D, \tag{5.7}$$

We proceed using the MLE estimator for both parameters λ and σ, which implies using iterative methods again as stated in [19], giving the estimators:

$$(\widehat{\lambda}_0, \widehat{\sigma}_0) = \arg\min_{\lambda, \sigma} \mathcal{C}(\lambda, \sigma; \overline{\mathcal{D}}). \tag{5.8}$$

It can be noted straight away that the value of $\widehat{\lambda}_0$ will be discarded since it was computed from a model that goes in direct contradiction with the assumptions we made in 5.1.3. However, $\widehat{\sigma}_0$ is the starting point of our heuristic which can be described in two steps. At the $r-$th iteration with $r \in [1 : R]$:

1. Minimise 5.6 using the value of $\widehat{\sigma}_{r-1}$ to obtain:

$$(\widehat{L}, \widehat{\boldsymbol{\eta}}) = \arg\min_{L, \boldsymbol{\eta}} \mathcal{C}(L, \boldsymbol{\eta}, \widehat{\boldsymbol{\lambda}}_{MLE}(L, \boldsymbol{\eta}), \widehat{\sigma}_{r-1}; \overline{\mathcal{D}}). \tag{5.9}$$

The change-points $\widehat{L}$ and the associated locations $\widehat{\boldsymbol{\eta}}$ are obtained by applying the PELT procedure [52], which improves the optimal partitioning approach through a lower, linear complexity. PELT is run several times on a penalty grid $(\beta_0, \ldots, \beta_q, \ldots, \beta_Q)$ where $\beta_0 < \cdots < \beta_q < \cdots < \beta_Q$. We obtain Q (not necessarily unique) segmentations of $\overline{\mathcal{D}}$. Eventually, the optimal penalty value for this step is selected using an elbow rule heuristic as proposed in [**?**]: segmentation scores are plotted against their corresponding number of change-points L. One looks for the number of breaks $\hat{L}$ that minimizes the sums of squares of two linear models respectively fitted on the $L \geq \hat{L}$ and the $L \leq \hat{L}$. We define $(\widehat{L}_r, \widehat{\eta}_r, \widehat{\boldsymbol{\lambda}}^r_{MLE}(\widehat{L}_r, \widehat{\eta}_r))$ the estimators associated to the segmentation selected with the elbow heuristic.

2. Compute $\widehat{\sigma}_r$. One can write the new negative log-likelihood of the whole signal $\overline{\mathcal{D}}$ defined by $(\widehat{L}_r, \widehat{\eta}_r, \widehat{\boldsymbol{\lambda}}^r_{MLE}(\widehat{L}_r, \widehat{\eta}_r))$ as a function of σ as follows:

$$g(\sigma; \overline{\mathcal{D}}) = \sum_{l=1}^{\widehat{L}_r+1} -\ln \mathcal{L}(\widehat{\lambda}_{r,l}, \sigma; \overline{y}_{\widehat{\eta}_{r,l-1}+1}, \ldots, \overline{y}_{\widehat{\eta}_{r,l}}), \tag{5.10}$$

where $\widehat{\lambda}_{r,l}$ is the $l-$th estimator in $\widehat{\boldsymbol{\lambda}}^r_{MLE}(\widehat{L}_r, \widehat{\eta}_r))$. $\widehat{\sigma}_r$ is found by minimizing 5.10 which can be perform using iterative procedures.

Simulation studies are available in Appendix B.1 showing the convergence of this heuristic. Once a suitable value of σ is obtained, the choice of the final penalty term β_K is driven by the CROPS algorithm [39], which computes all optimal segmentations as the penalty varies over some interval. Eventually, the final penalty value is selected using an elbow rule heuristic.

5.2.2. Spatial clustering

In any of stationary intervals identified in the previous step, the measurements are assumed to be consistent with the homogeneity assumptions represented by the graph $G = (V, E)$. A natural way of assessing the actual regularity of the measurements would be to use graph signal processing techniques see e.g. [76, 90]. However the irregular, unaligned, sparse and censored nature of the measurements at each station, prevents the use of such methods. The measurements are also incompatible with techniques designed to detect anomalous clusters in a graph see for instance [5].

To circumvent this problem, we propose to leverage the graphical representation to build spatial aggregates and to assess homogeneity at this aggregated level. This corresponds to clustering the stations using the graph structure. Nodes of each connected component $(\mathcal{K}_1, ..., \mathcal{K}_P)$ of the graph $G = (V, E)$ are clustered using a Ward hierarchical clustering method implemented on the shortest path distance computed from the edge weights.

The goal is to successfully create a global partition in M cluster of stations in the presence of P non connected components in the graph. This raises the question of how to dispatch these M clusters among the non connected components. We have developed two methods. The first one proceed in a greedy way, the second is based on dynamic programming. Both ot them are

based the standard definition of inertia given for the clustering $\mathcal{P} = (C_1, \ldots, C_M)$ by

$$W(\mathcal{P}) = \sum_{m=1}^{M} \frac{1}{|C_m|} \sum_{v_i, v_j \in C_k} d_{ij}^2, \tag{5.11}$$

where d_{ij}^2 is the square of the shortest path distance in G between vertices v_i and v_j, and $|A|$ denotes the cardinality of set A. Clustering with a small inertia contain clusters that group close monitoring stations according to the graph G.

1. **The greedy clustering method:** the initial global clustering of V is obtained by assigning all vertices in a connected component to the same cluster. Subsequent levels of the global hierarchy are obtained by replacing the clusters of a connected component by the next refined level of the local hierarchy. At each step of the refinement, we select the component that reduce the most the inertia of the clustering 5.11.

2. **A clustering method based on dynamic programming:** this approach is derived from [41]. This paper shows that is possible to create a partition of the stations graph into P components and to perform a segmentation of each the P components using a total number M of segments. The M segments are distributed among the P components in an optimal way using dynamic programming. Our context is a little bit simpler than [41] since the P components are already known and doesn't have to be estimated.

The algorithms for both methods are provided in Appendix B.2. To select the final clustering in the hierarchy, we use the same decision rule as 5.2.1. This time, the inertia of the clustering is plotted against the corresponding number of clusters M. We look for the number of breaks M^* that minimizes the sums of squares of two linear models respectively fitted on the $M \geq M^*$ and the $M \leq M^*$.
Notice that we rely on a simple graph clustering approach for two main reasons. Firstly, we do not expect graphs of monitoring stations to exhibit the specific characteristics of complex networks (such as very high degree vertices, small diameter, etc. see e.g. [72]) that justify the use of techniques such as maximal modularity clustering see e.g. [30]. On the contrary, simpler approaches that interpret shortest paths weights as dissimilarities should be sufficient see e.g. [86]. Secondly, we work on relatively small graphs with even smaller connected components and we do not face computational issues associated to hierarchical clustering. Finally, it is important to note that the spatial clustering is independent from the temporal context $[\widehat{\eta}_l, \widehat{\eta}_{l+1}]$. The clustering is performed on the graph $G = (V, E)$ composed of all stations available in the data.

5.2.3. Anomaly detection

Two types of anomalous clusters are targeted: either clusters with anomalous stations, or wholly anomalous clusters. Clusters containing anomalous stations are detected by studying the homogeneity of the measurements provided by the stations in a given spatial cluster. Anomalous clusters of stations are detected by simply pooling all measurements of each cluster to estimate

the local use of the substance and detect large rates. We derive in this section two anomaly scores covering those cases.

For the first case, we need to assess the homogeneity of the measurements of the stations in a spatial cluster for a stationary time interval. As pointed out previously, the number of measurements provided by a single station is usually quite small, especially when we consider a single stationary interval. As a consequence classical distances between empirical distributions are not appropriate, mainly because the measurements of two stations do not have any value in common. Then the Kolmogorov-Smirnov statistics will be essentially driven by the number of observed values rather than the actual values, while other quantities, such as the Jensen-Shannon divergence, cannot be properly estimated (see appendix B.3). For this reason, we propose to use the Wasserstein w_1 distance [96] adapted for left censored variables. For two discrete distributions on $\mathbb{R}$, it is expressed as the L^1-distance between their cumulative distribution functions and is therefore simple to compute.

The measurement homogeneity of the clusters obtained in Section 5.2.2 is therefore defined as the mean within cluster empirical Wasserstein average distance of a station measurements to the others. Denoting C_m the m-th cluster and $|C_k|$ the number of stations present in C_m, $w_1(\mathbf{y}_i, \mathbf{y}_j)$ the empirical 1-Wasserstein distance between the data of stations v_i and v_j, this quantity is expressed as

$$\bar{W}_k = \frac{1}{|C_m|(|C_m|-1)} \sum_{1 \leq j \leq |C_m|} \sum_{1 \leq i \leq |C_m|, i \neq j} w_1(\mathbf{y}_i, \mathbf{y}_j). \tag{5.12}$$

The second type of potentially anomalous clusters are simply associated to the presence of quantified measurements and high values of concentration. Thus we estimate for each spatial cluster C_m the parameters of distribution Q (see Section 5.1.3) on the pooled measurements obtained from all the stations of the cluster during the chosen stationary interval. From those parameters, we compute a statistics, denoted $\bar{I}_m$, used as a proxy for the intensity of the measurements (see Section 5.4.3 for an example). Hence we consider a low concentration to be the normal case, but we do not define a threshold between normal clusters and abnormal ones.

Each cluster C_m is therefore characterised by two values $(\bar{W}_m, \bar{I}_m)$. To select potentially anomalous clusters, we use a multi-objective optimisation approach, considering that both characteristics are equally interesting. Following [51], we say that $X_k = (\bar{W}_m, \bar{I}_m)$ is *Pareto dominated by* $X_l = (\bar{W}_l, \bar{I}_l)$, and we write $X_m \prec X_l$ if and only if

$$\left((\bar{W}_m < \bar{W}_l) \text{ and } (\bar{I}_m \leq \bar{I}_l)\right) \text{ or } \left((\bar{W}_m \leq \bar{W}_l) \text{ and } (\bar{I}_m < \bar{I}_l)\right).$$

The level 1 Pareto optimal front is the set of maximal points for $\prec$. Level b with $b > 1$ is defined recursively as the optimal Pareto front computed for the set of points that do not belong to the optimal Pareto front of levels $1, \ldots, b-1$. Therefore clusters in the level 1 Pareto front are remarkable in the sense that there is no other cluster with higher heterogeneity and more extreme measurements. We define these clusters as anomalous. Pareto front and levels are evaluated using the Skyline algorithm [10, 25].

5.3. Data presentation

The methodology introduced in the above sections will be illustrated next using a case study on the prosulfocarb concentration [71] in Centre-Val de Loire. This chemical compound is mainly used as a herbicide in field crops, with a typical period of active use in autumn. The monitoring of its concentrations in surface waters has been subject to increasing attention due to its aquatic ecotoxicology [3, 2].

5.3.1. Time period and geographical area selection

Prosulfocarb usage was banned in France before 2007. A market re-authorisation was issued by the French Observatory on Pesticide Residues (now part of the ANSES[2]) in 2009. Since then, two modifications of the authorisation for use have been put in place, in November 2018 and in November 2019 respectively. Both changes consist in restrictions of use, one imposing specific equipment for application, the other restricting the application schedule in the presence of non-target crops next to the treated area. Motivated by these changes in regulation, the time period chosen for our study spans from January 1, 2007, to April 8, 2022. Moreover, our study focuses on the geographical area of French Centre-Val de Loire region. Indeed, between 2009 and today, the annual mass of prosulfocarb sold in this region exploded, making it rise from the 17th most sold substance in 2009 to the 4th in 2017 (see Figure B.6 in Appendix B.4.2). This region is also characterised by high concentrations of prosulfocarb target crops (such as the Beauce plains) (see Figure B.5 in Appendix B.4.1). Many target crops (cereal crops) are also concentrated in the region. These two elements combined guarantee a significant use of the product in this area. Thus, we expect significant variations in concentration values in this area during this period.

Data about surface water quality in France is available from the French Biodiversity Agency [74]. We collected from the site the data selected above[3]. These choices led to a data set $\mathcal{D}$ comprising 420 monitoring stations that performed 14,203 measurements. Each measurement is described by the monitoring station ID, the sampling date, the quantification limit (LOQ), and the concentration measurement value, if the concentration exceeds the LOQ. In the data used in this work, the LOD is unknown: the left censoring phenomenon corresponds therefore to the LOQ of the measuring stations. When both limits are known, one can adapt the model proposed in Section 5.1.2 to take both of them into account: this would translate into a slightly more complex likelihood as the one derived in Chapter 4 as we need to consider three cases (when the concentration is between 0 and the LOD, when the concentration is between the LOD and the LOQ, and finally when the concentration is observed and larger than the LOQ). Among the 14,203 recorded measurements during the period of interest, only 14.11% were above the quantification limit. Figure 5.1 shows the distribution of the number of measurements per

[2]ANSES stands for *Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail*, i.e., French Agency for Food, Environmental and Occupational Health & Safety.

[3]Data exported in September 2020 using `http://www.naiades.eaufrance.fr/acces-donnees#/physicochimie/resultats?debut=09-01-2007&fin=08-09-2020&regions=24&parametres=1092&fractions=23&supports=3&qualifications=1`

station: the mean (rounded to the closest integer) and median number of samples collected by each monitoring station are respectively 34 and 19. This illustrates that sampling rates are different across stations, most of them making few measures, and the monitoring process is heterogeneous.

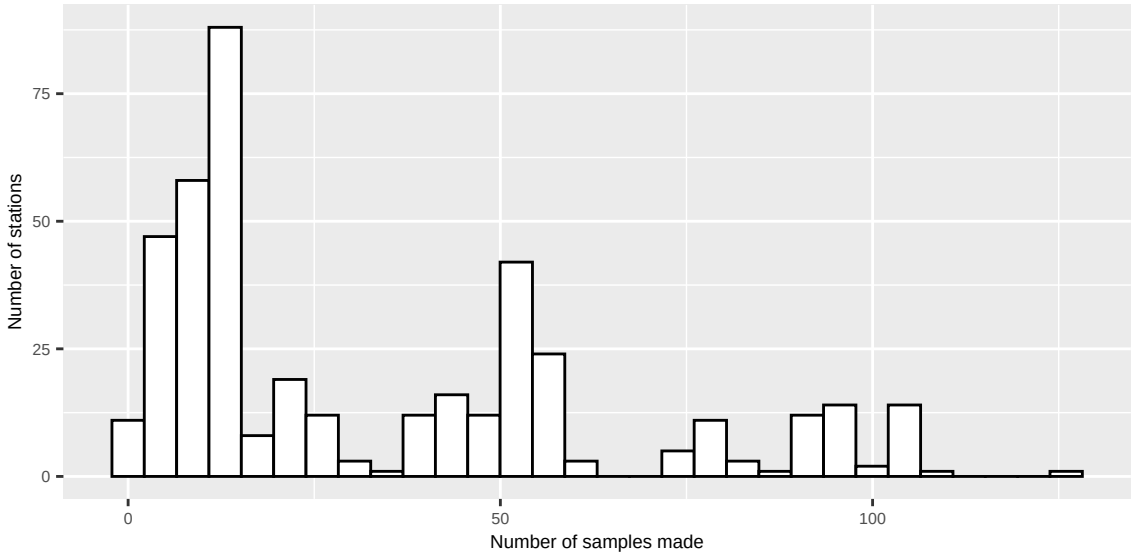

Figure 5.1: Distribution of the number of measurements per station.

The coarse representation $\overline{\mathcal{D}}$ of the monitoring data $\mathcal{D}$ is obtained by computing the maximum daily values across the available stations. This yields the time series illustrated in Figure 5.2. The aggregated series contains 2,150 values, among which 22.51% are quantified.
One may note here that despite the aggregation process, the coarse series remains irregularly sampled, and that for about two thirds of the days included in the studied time span, no measurements were made.

5.3.2. Graphical representation of the station network

The stations network $G = (V, E)$ introduced in Section 5.1.1 is built using the hydrographic map of the Centre-Val de Loire region. Indeed, once the monitoring stations are geo-localized through their GPS coordinates, one still has to compute the edges between them, as well as the associated weights.
For the data at hand, edges are determined using the river network. A database provided by the French National Institute of Geographic and Forest Information (IGN) [46] contains a fine-grained description of rivers, encoded as sequences of hydrographic sections (or river sections). River sections are segments with constant geographic and hydrographic attributes.
The procedure used for computing the edges in the stations network based on the river network may be summarised as follows:

1. One starts by building a river network $R = (S, H)$, where the vertices S are made of the connecting points between the river sections, and the edges H contain all sections.

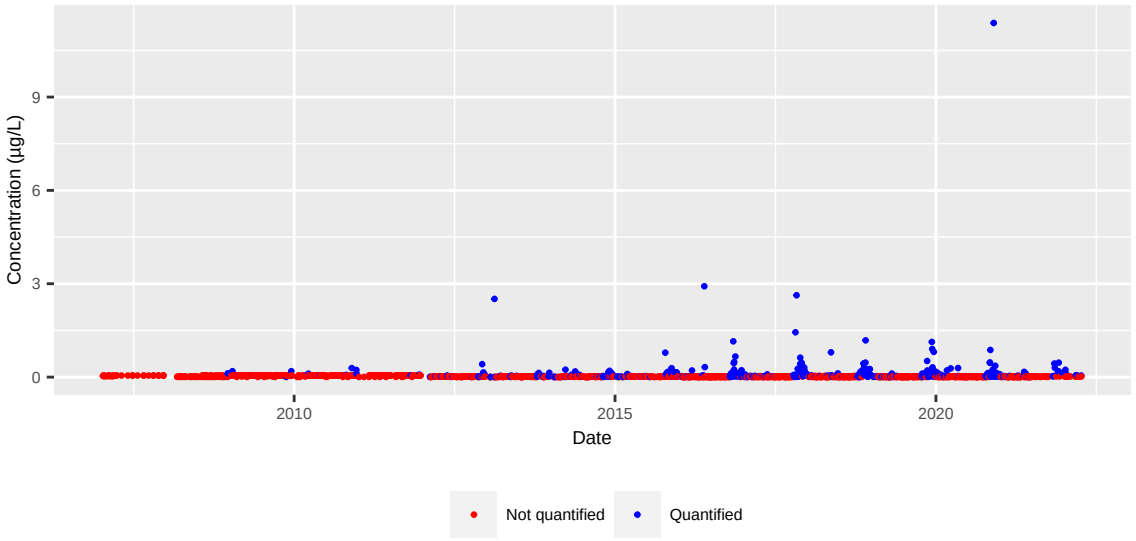

Figure 5.2: Plot of daily maximum concentrations

Each edge is thus naturally weighted by the length (in meters) of the corresponding river section.

2. Each monitoring station v_i in V is assigned to the closest node $\tilde{s}_i$ in the river network R, by minimizing the geographical distance between the station v_i and all connecting points

$$\tilde{s}_i = \min_{s \in S} d(v_i, s).$$

3. Given two stations $v_i, v_j \in V$ and their associated connecting points $\tilde{s}_i, \tilde{s}_j \in S$, an edge will be generated between v_i and v_j if there exists at least one path between $\tilde{s}_i$ and $\tilde{s}_j$. Furthermore, the weight associated to an edge (v_i, v_j) is equal to the length of the shortest path between $\tilde{s}_i$ and $\tilde{s}_j$.

One may notice at this point that the above procedure may result into an unconnected graph, with several connected components. For illustration, Figure 5.3 displays the graph of all stations that made at least one sample during the obsetvation period. It is not fully connected and exhibits 9 distinct connected components.

5.4. Results

5.4.1. Temporal segmentation

First, the coarse-grained time series $\overline{\mathcal{D}}$ in Figure 5.2 is segmented using the change-point detection procedure described in Section 5.2.1. The results of the heuristic proposed to estimate σ are shown in Figure 5.4.

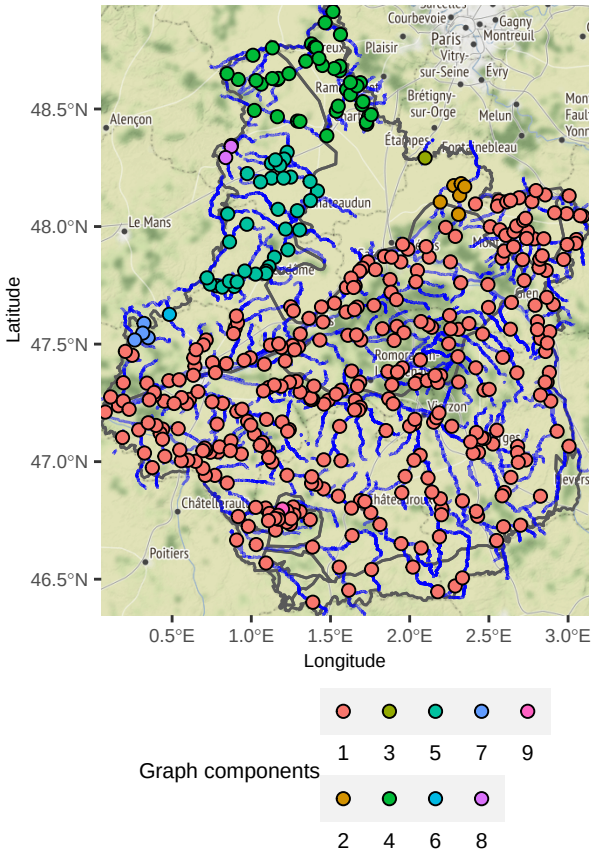

Figure 5.3: Map of the non connex components in the station graph.

The penalty grid $[\frac{\ln K}{4}, 4 \ln K]$ was chosen for the heuristic where K is the number of daily maximum concentrations available, here $K = 2150$.. This interval allowed to obtain various segmentation results with the number of change points varying from 1 to 30. With such kind of range in the results, it ensured a central position of the elbow in the plot of the cost of these segmentations against their respective number of change-points. The precision stopping criterion was set to $10^{-}3$. From the application point of view, the assumption that σ is a fixed parameter throughout the series $\overline{\mathcal{D}}$ corresponds to the hypothesis that the differences in usage and diffusion of the prosulfocarb among the different users is captured by the shape parameter, and should not vary much over time. On the contrary, the overall average usage of prosulfocarb varies, and this dependency is captured by changes in the rate parameter. Hence, after computing the MLE of σ, $\hat{\sigma}_{MLE}$, over the whole time series, change-points and rate parameters over each temporal segment are estimated by minimizing the cost function in Equation 5.5. Let us remark here that the estimated value of the shape parameter is $\hat{\sigma}_{MLE} = 0.4$. This confirms the data has a heavier tail than an exponential distribution (σ=1), and that the assumption of using Weibull distributions for our data is appropriate.

In the change-point detection procedure, the penalty value for the PELT algorithm was calibrated using a large range of values explored according to the CROPS algorithm. The range, inspired by the BIC criterion, was set to $[\frac{\log(K)}{5}, 5 \log(K)]$. Note that when using the BIC penalty in change point detection, the penalty term written in section 5.2.1 becomes :

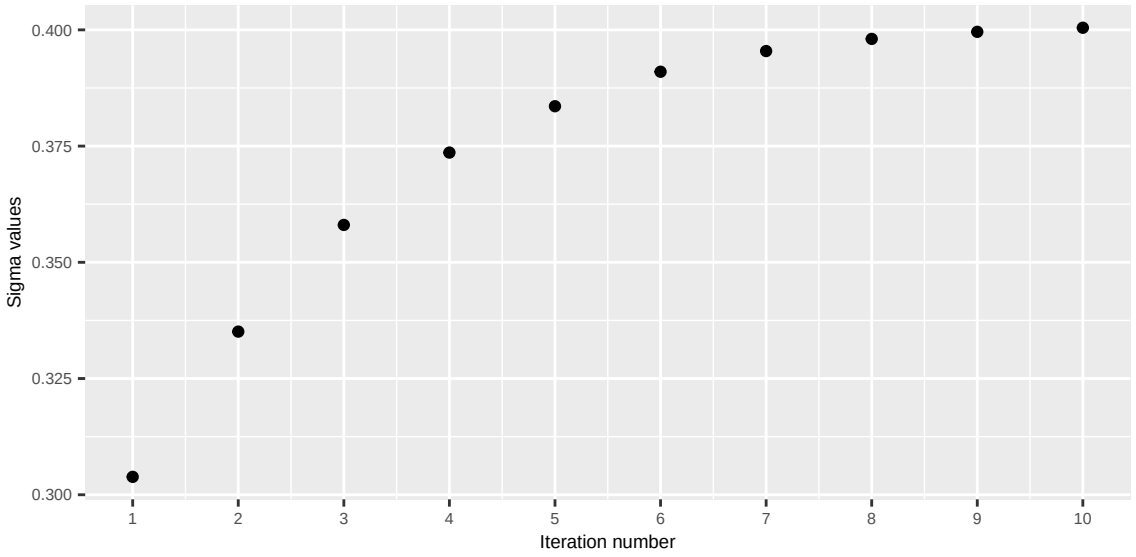

Figure 5.4: Plot of successive $\widehat{\sigma}$ values. We stopped the to iterate when the $|\widehat{\sigma}_b - \widehat{\sigma}_b| \leq 10^{-3}$

$\beta_K(L+1)D = \frac{D}{2}\log(K)(L+1) = \frac{1}{2}\log(K)(L+1)$. The range chosen allows to screen an interval of penalties containing the BIC penalty.
The penalty calibration procedure resulted in 15 different segmentations, with a number of change-points ranging from 1 to 30. The best segmentation is selected using the elbow method, as illustrated in Figure B.7 in Appendix B.4.3. This amounts to a temporal segmentation with $\hat{L} = 13$ change-points, illustrated in Figure 5.5.
According to Figure 5.5, the usage of prosulfocarb in Centre-Val de Loire shows different patterns throughout time. Before 2016, most of the values are not quantified, and there are almost no change-points detected. Starting with 2016, two regimes of pesticide usage appear to emerge, and correspond respectively to the periods of intensive usage of prosulfocarb and to the off-peak periods. Indeed, the starting dates of the peak periods coincide with the season where the substance is spread, which is Autumn. The emergence of this two-regime pattern, alternating high concentration values during the peak periods and low concentration values during the off-peaks, is correlated with an important increase in the prosulfocarb sales as shown in Figure B.6 in Appendix B.4.2.

5.4.2. Spatial segmentation

The second step of the analysis consists in the spatial segmentation using the graph-based clustering on the monitoring stations network. This step is strictly independent from the temporal segmentation.
During the whole observation period, 420 monitoring stations only produced at least one measure. The spatial clustering algorithm was applied with a number of potential clusters varying between 7 and 35. The minimum number of clusters is equal to the number of connected components in the graph composed of more than one station plus one cluster. There are 6 components with more than one station and we add a supplementary cluster at the initialisation of the clus-

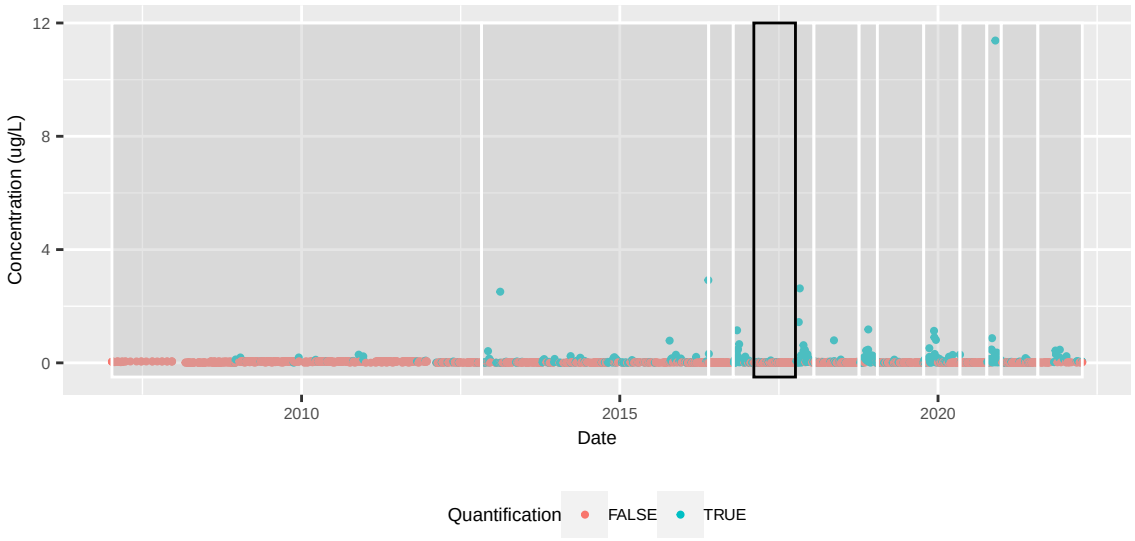

Figure 5.5: Best segmentation found by the change-point detection procedure with CROPS-based penalty tuning. The dates of the breaks are : October 20, 2012; May 25, 2016; October 13, 2016; February 7, 2017; October 5, 2017; January 19, 2018; October 5, 2018; January 18, 2019; October 11, 2019; May 6, 2020; October 7, 2020; December 20, 2020; July 27, 2021.The black rectangle corresponds to the selected temporal segment in section 5.4.2

tering procedure. The optimal number of clusters was selected using the elbow method applied to the inertia curve. According to this criterion, illustrated in Figure in Appendix B.4.3, the best solution is made of a 15-clusters configuration. The spatial segmentation is illustrated in Figure 5.6. The algorithm based on dynamic programming 6 was used to create the partition of the station graph. To check the relevance of the homogeneity assumption formulated in section 5.2.2, let us focus on a specific temporal segment. An off-peak period, spanning between February 8, 2017 and October 4, 2017 was selected. This period was identified as a homogeneous temporal segment by the change-point detection procedure. This period is highlighted by the black rectangle in Figure 5.5. We proceeded in two steps. First we pooled all samples made during that specific period of time. From all those samples, we can identified the active clusters during that period of time, there were 13 out of 15. Then, for all active clusters, we computed the within average empirical pairwise Wasserstein distance of the active stations of a cluster and observe that for 10 clusters out of 13, this indicator is less than 0.0015, whereas the global average pairwise Wasserstein distance for the 149 stations is 0.003. This suggests that the distance chosen for our station graph is indeed a good proxy of the homogeneity in the concentration space. Additional comments can be made when we look at the geography of the region. Some clusters are overlapping with hydro-ecoregions. Hydro-ecoregions are geographic entities in which hydrographic ecosystems share common characteristics. The criteria defining them combine properties of geology, terrain and climate [98]. The borders of those regions are drawn in grey in Figure . This ensures that the substances will have homogeneous dispersion properties on these clusters (see clusters 7). As expected the biggest component in Figure 5.3 is the most segmented. Some clusters are easy to identify, for instance clusters 12 corresponds to

the Indre river. Cluster 10 is identified as the most western part of the Loire and its tributaries mainly the Vienne and the Creuse rivers. Clusters 1,7,9 and 10 are a little bit harder to identify. If one look closely at the map of the region, there is a high presence of small channels all across this part of the region.

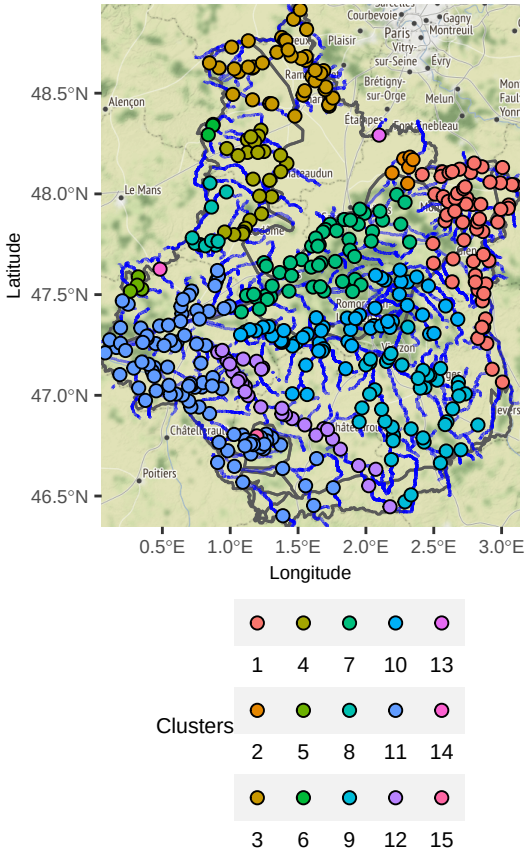

Figure 5.6: Map of geographical clusters.

5.4.3. Anomalous cluster identification

We now focus on locating spatial patterns during time segments identified in section 5.4.1. Peak periods in prosulfocarb use are less likely to produce rich spatial patterns since they correspond to a intensive overall use in the region. This why an off-peak period was investigated instead. In the rest of this case study, we selected the segment highlighted in black in figure 5.5. It is delimited by the dates February 8, 2017 and October 4, 2017. This introduces a context to the anomaly detection: a global non use of the substance.
Following the methodology proposed in 5.2.3, the scaling parameter λ_k of the aggregated data of each spatial cluster found in Section 5.4.2 was estimated. The statistics $\bar{I}_k$ was set to $1/\hat{\lambda}_k$. The Pareto front involving the two descriptors $\bar{W}_k$ and $\bar{I}_k$ was computed. It led to the cluster ranking displayed in Figure 5.7 using the *rPref* package [83]. We recall that the selected time segment corresponds to a period of non-use of prosulfocarb. From this it can be deduced that

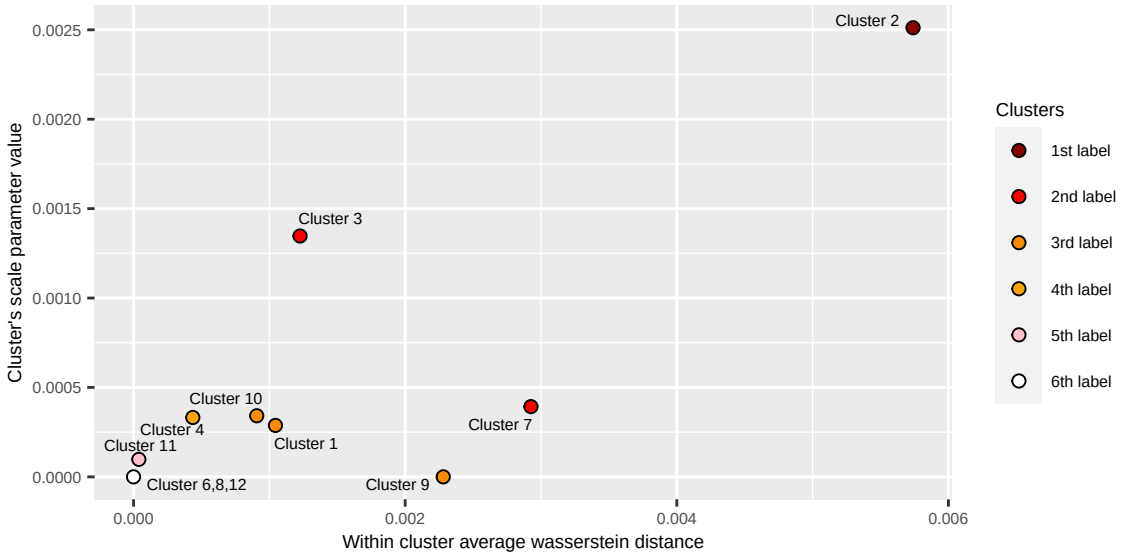

Figure 5.7: Clusters pareto front.

finding quantified measurements of the substance during this period is an anomaly. Three clusters stood out with a Pareto front levels of 1 and 2. Among them we can find on Figure 5.7:

- **Cluster 2:** which is the most anomalous cluster. There is a bias coming from the number of samples made during that time period. Only 11 measures were reported. However, it is interesting to note that this cluster has a 27.27% rate of quantification which corresponds to 3 quantified measurements. The rate of quantification has a huge influence on the estimated scale parameter of the cluster. It is then logical to find this cluster dominating the other on this axis. This cluster didn't record the maximum concentration during the period but its highest quantification value is up to 0.031 μg/L which the third highest value recorded in the temporal segment. Combined with the high quantification rate, it implies that the mean within Wasserstein distance is elevated.

- **Clusters 3 and 7:** which are Pareto level 2 clusters. Cluster 3 has a 6.09% quantification rate which higher than cluster 7 (4.48%). This explains its higher position on the scale parameter estimate axis. Its maximum value is 0.039 μg/L which is smaller than the maximum in cluster 7 which is 0.087 μg/L. The difference in within Wasserstein distance is higher in cluster 7 because it has a station that made a very high quantification compare to other stations. the recorded 0.087 μg/L is actually the maximum of concentration of the whole temporal segment.

It is interesting to note that the Pareto front level is not uniformly distributed in the region. The three anomalous clusters are located in the north and east of the region. It could be related to the agricultural practices and land use. For the sake of the argument, we present in Appendix B.4.1 the map of barley and wheat crops in Centre-Val de Loire. In future works, we shall investigate the spatial correlation between anomalous clusters and areas with high concentration of these crops. Figure displays the Pareto front levels on the station map 5.8.

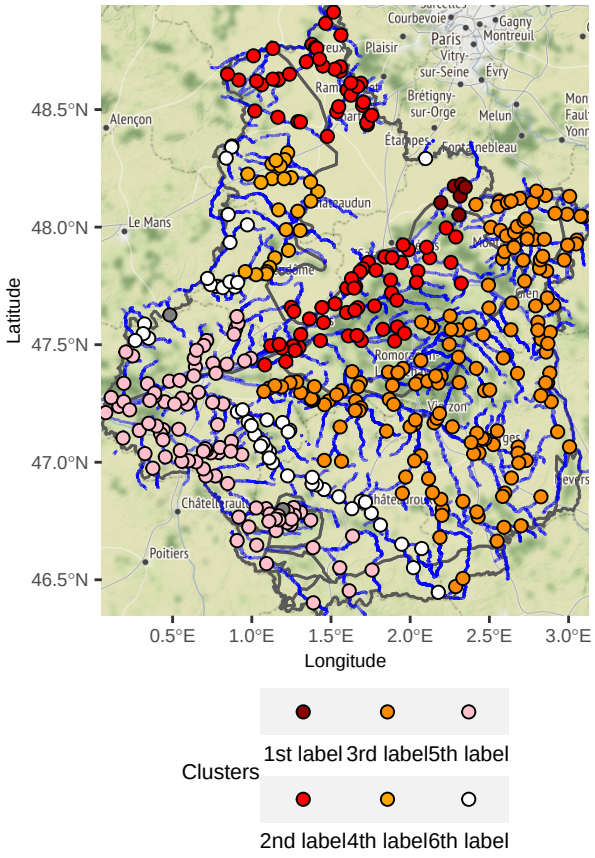

Figure 5.8: Mapped pareto front.

6. Conclusion

6.1. Summary

6.2. Openings

In this work we have seen that it is possible to extract information from data whose properties make modelling difficult. Further work is conceivable in the future. First, we have never addressed the issue of simultaneous monitoring of multiple substances. This is possible with the multivariate change point detection methods presented in Chapter 2, but these methods will always depend on heterogeneous spatiotemporal sampling. It can also be added that if different segmentations are obtained for different substances, there are ways to compare the positioning of these breaks. This topic is addressed, for example, in [**?**].
Another starting point for future work is to try to optimise the placement of the stations and their sampling frequencies. It is obvious that a synchronous sampling rate of all stations would allow to observe the dynamics of the dispersion of the substance in space and time (in areas where there is only one emission source for the substance). Optimising the placement and sampling frequency of the stations is a similar issue to optimal design.

Bibliography

[1] Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

[2] Agriculture & Environment Research Unit (AERU) at the University of Hertfordshire (2021). Prosulfocarb (Ref: SC 0574). `https://sitem.herts.ac.uk/aeru/ppdb/en/Reports/557.htm`. Retrieved: March 1, 2022. Part of [55].

[3] ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) (2018). Prosulfocarbe (phytopharmacovigilance). `https://www.anses.fr/fr/system/files/Fiche_PPV_Prosulfocarbe.pdf`. Retrieved: March 1, 2022 (French document).

[4] Antweiler, R. C. and Taylor, H. E. (2008). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. summary statistics. *Environmental Science & Technology*, 42(10):3732–3738.

[5] Arias-Castro, E., Candès, E. J., and Durand, A. (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278 – 304.

[6] Aznar, R., Moreno-Ramón, H., Albero, B., Sánchez-Brunete, C., and Tadeo, J. L. (2017). Spatio-temporal distribution of pyrethroids in soil in mediterranean paddy fields. *Journal of Soils and Sediments*, 17(5):1503–1513.

[7] Bai, J. (1994). LEAST SQUARES ESTIMATION OF a SHIFT IN LINEAR PROCESSES. *Journal of Time Series Analysis*, 15(5):453–472.

[8] Bardet, Jean-Marc, Brault, Vincent, Dachian, Serguei, Enikeeva, Farida, and Saussereau, Bruno (2020). Change-point detection, segmentation, and related topics. *ESAIM: ProcS*, 68:97–122.

[9] Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Change: Theory and Application*, volume 15. prentice Hall Englewood Cliffs.

[10] Borzsony, S., Kossmann, D., and Stocker, K. (2001). The skyline operator. In *Proceedings 17th International Conference on Data Engineering*, pages 421–430.

[11] Bouveyron, C., Jacques, J., Schmutz, A., Simoes, F., and Bottini, S. (2022). Co-clustering of multivariate functional data for the analysis of air pollution in the south of france. *Annals of Applied Statistics*, 16(3):1400–1422.

[12] Bunce, C., Carr, J. R., Nienow, P. W., Ross, N., and Killick, R. (2018). Ice front change of marine-terminating outlet glaciers in northwest and southeast greenland during the 21st century. *Journal of Glaciology*, 64(246):523–535.

[13] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

[14] Canales, R. A., Wilson, A. M., Pearce-Walker, J. I., Verhougstraete, M. P., and Reynolds, K. A. (2018). Methods for handling left-censored data in quantitative microbial risk assessment. *Applied and Environmental Microbiology*, 84(20):e01203–18.

[15] Ccanccapa, A., Masiá, A., Andreu, V., and Picó, Y. (2016). Spatio-temporal patterns of pesticide residues in the turia and júcar rivers (spain). *Science of The Total Environment*, 540:200–210. 5th Special Issue SCARCE: River Conservation under Multiple stressors: Integration of ecological status, pollution and hydrological variability.

[16] Chen, J. and Gupta, A. K. (2012). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance*. Springer.

[17] Chen, J., Kim, S.-H., and Xie, Y. (2020). S3t: A score statistic for spatiotemporal change point detection. *Sequential Analysis*, 39(4):563–592.

[18] Chen, S., Gopalakrishnan, P., et al. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA broadcast news transcription and understanding workshop*, volume 8, pages 127–132. Citeseer.

[19] Cohen, A. C. (1965). Maximum likelihood estimation in the weibull distribution based on complete and on censored samples. *Technometrics*, 7(4):579–588.

[20] Costa, M., Gonçalves, A. M., and Teixeira, L. (2016). Change-point detection in environmental time series based on the informational approach. *Electronic Journal of Applied Statistical Analysis*, 9(2):267–296.

[21] Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

[22] Croghan, W. and Egeghy, P. P. (2003). Methods of dealing with values below the limit of detection using sas carry. In *The Proceedings of the SouthEast SAS Users Group*.

[23] Delignette-Muller, M. L. and Dutang, C. (2015). fitdistrplus: An r package for fitting distributions. *Journal of statistical software*, 64:1–34.

[24] Einmahl, J. H. J. and McKeague, I. W. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2):267–290.

[25] Endres, M., Roocks, P., and Kießling, W. (2015). Scalagon: an efficient skyline algorithm for all seasons. In *International Conference on Database Systems for Advanced Applications*, pages 292–308. Springer.

[26] European Food Safety Authority (2010). Management of left-censored data in dietary exposure assessment of chemical substances. *EFSA Journal*, 8(3):1557.

[27] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

[28] Fearnhead, P., Maidstone, R., and Letchford, A. (2018). Detecting changes in slope with an l_0 penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275.

[29] Figueiredo, D. M., Duyzer, J., Huss, A., Krop, E. J., Gerritsen-Ebben, M., Gooijer, Y., and Vermeulen, R. C. (2021). Spatio-temporal variation of outdoor and indoor pesticide air concentrations in homes near agricultural fields. *Atmospheric Environment*, 262:118612.

[30] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174.

[31] Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(3):495–580.

[32] Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6).

[33] Gillaizeau, F., Gal, C. L., Maudet, C., Fournier, M., and Leuillet, S. (2020). Méthodes de gestion des valeurs sous des seuils de détection ou de quantification. *Revue d'Épidémiologie et de Santé Publique*, 68:S137.

[34] Gillespie, B. W., Chen, Q., Reichert, H., Franzblau, A., Hedgeman, E., Lepkowski, J., Adriaens, P., Demond, A., Luksemburg, W., and Garabrant, D. H. (2010). Estimating population distributions when some data are below a limit of detection by using a reverse kaplan-meier estimator. *Epidemiology*, 21(4):S64–S70.

[35] Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493.

[36] Harchaoui, Z., Moulines, E., and Bach, F. (2008). Kernel change-point analysis. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

[37] Harvell, C. D., Kim, K., Burkholder, J. M., Colwell, R. R., Epstein, P. R., Grimes, D. J., Hofmann, E. E., Lipp, E. K., Osterhaus, A. D. M. E., Overstreet, R. M., Porter, J. W., Smith, G. W., and Vasta, G. R. (1999). Emerging marine diseases–climate links and anthropogenic factors. *Science*, 285(5433):1505–1510.

[38] Haynes, K., Eckley, I. A., and Fearnhead, P. (2014). Efficient penalty search for multiple changepoint problems. *arXiv preprint arXiv:1412.3617*.

[39] Haynes, K., Eckley, I. A., and Fearnhead, P. (2017). Computationally efficient change-point detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143.

[40] He, H. and Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, 16(3).

[41] Hébrail, G., Hugueney, B., Lechevallier, Y., and Rossi, F. (2010). Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*, 73(7-9):1125–1141.

[42] Hewett, P. and Ganser, G. H. (2007). A comparison of several methods for analyzing censored data. *The Annals of occupational hygiene*, 51 7:611–32.

[43] Höhle, M. (2010). Online change-point detection in categorical time series. In *Statistical modelling and regression structures*, pages 377–397. Springer.

[44] Höök, M. and Tang, X. (2013). Depletion of fossil fuels and anthropogenic climate change—a review. *Energy Policy*, 52:797–809.

[45] Institut National de l'Information Géographique et Forestière (2020). Registre parcellaire graphique (RPG). `https://geoservices.ign.fr/bdtopo`. Retrieved: March 1, 2022.

[46] Institut National de l'Information Géographique et Forestière (2021). BD TOPO®. `https://geoservices.ign.fr/bdtopo`. Retrieved: March 1, 2022.

[47] Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., San, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108.

[48] Jensen, R. E. (1969). A dynamic programming algorithm for cluster analysis. *Operations Research*, 17(6):1034–1057.

[49] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*. Wiley-Interscience.

[50] Khopkar, S. (2007). *Environmental pollution monitoring and control*. New Age International.

[51] Kießling, W. (2002). Chapter 28 - foundations of preferences in database systems. In Bernstein, P. A., Ioannidis, Y. E., Ramakrishnan, R., and Papadias, D., editors, *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases*, pages 311–322. Morgan Kaufmann, San Francisco.

[52] Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

[53] Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1):79–102.

[54] Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, 3:637–662.

[55] Lewis, K., Tzilivakis, J., Warner, D., and Green, A. (2016). An international database for pesticide risk assessments and management. *Human and Ecological Risk Assessment: An International Journal*, 22(4):1050–1064.

[56] Li, S., Xie, Y., Dai, H., and Song, L. (2015). M-statistic for kernel change-point detection. *Advances in Neural Information Processing Systems*, 28.

[57] Li, W., Guo, W., Luo, X., and Li, X. (2010). On sliding window based change point detection for hybrid SIP DoS attack. In *2010 IEEE Asia-Pacific Services Computing Conference*. IEEE.

[58] Li, Y., Bao, T., Shu, X., Gao, Z., Gong, J., and Zhang, K. (2021). Data-driven crack behavior anomaly identification method for concrete dams in long-term service using offline and online change point detection. *Journal of Civil Structural Health Monitoring*, 11(5):1449–1460.

[59] Liang, Z., Qian, S. S., Wu, S., Chen, H., Liu, Y., Yu, Y., and Yi, X. (2019). Using bayesian change point model to enhance understanding of the shifting nutrients-phytoplankton relationship. *Ecological Modelling*, 393:120–126.

[60] Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., and Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and ecological statistics*, 21(3):411–433.

[61] Liu, C., Chen, Y., Chen, F., Zhu, P., and Chen, L. (2022). Sliding window change point detection based dynamic network model inference framework for airport ground service process. *Knowledge-Based Systems*, 238:107701.

[62] Liu, S., Wright, A., and Hauskrecht, M. (2017). Change-point detection method for clinical decision support system rule monitoring. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 126–135. Springer.

[63] Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2011). Distributed detection/localization of change-points in high-dimensional network traffic data. *Statistics and Computing*, 22(2):485–496.

[64] Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, 156(4):133–162.

[65] Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2016). On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533.

[66] Majumdar, A., Gelfand, A. E., and Banerjee, S. (2005). Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference*, 130(1):149–166. Herman Chernoff: Eightieth Birthday Felicitation Volume.

[67] Manly, B. F. (2008). *Statistics for Environmental Science and Management*. Chapman and Hall/CRC.

[68] Marchant, C., Leiva, V., Christakos, G., and Cavieres, M. F. (2018). Monitoring urban environmental pollution by bivariate control charts: New methodology and case study in santiago, chile. *Environmetrics*, 30(5):e2551.

[69] Mitra, S. and Kundu, D. (2008). Analysis of left censored data from the generalized exponential distribution. *Journal of Statistical Computation and Simulation*, 78(7):669–679.

[70] Mori, A. S., Furukawa, T., and Sasaki, T. (2012). Response diversity determines the resilience of ecosystems to environmental change. *Biological Reviews*, 88(2):349–364.

[71] National Center for Biotechnology Information (2022). PubChem Compound Summary for CID 62020, Prosulfocarb. `https://pubchem.ncbi.nlm.nih.gov/compound/Prosulfocarb`. Retrieved: March 1, 2022.

[72] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

[73] Nougadère, A., Merlo, M., Héraud, F., Réty, J., Truchot, E., Vial, G., Cravedi, J.-P., and Leblanc, J.-C. (2014). How dietary risk assessment can guide risk management and food monitoring programmes: The approach and results of the french observatory on pesticide residues (anses/orp). *Food Control*, 41:32–48.

[74] Office français de la biodiversité. Naïades, données sur la qualité des eaux de surface. `http://www.naiades.eaufrance.fr/`, `http://www.ofb.gouv.fr/`. Retrieved: March 1, 2022.

[75] Office français de la biodiversité and Système d'Information sur l'Eau (2021). Achats de pesticides par code postal. `https://geo.data.gouv.fr/fr/datasets/bdc2c6f21f70acccfea73445f68a5f0d6ee5b7c1`, `https://www.eaufrance.fr/`, `http://www.ofb.gouv.fr/`. Retrieved: March 1, 2022.

[76] Ortega, A., Frossard, P., Kovačević, J., Moura, J. M. F., and Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828.

[77] Ozgul, A., Childs, D. Z., Oli, M. K., Armitage, K. B., Blumstein, D. T., Olson, L. E., Tuljapurkar, S., and Coulson, T. (2010). Coupled dynamics of body mass and population growth in response to environmental change. *Nature*, 466(7305):482–485.

[78] Pettitt, A. (1980). Some results on estimating a change-point using non-parametric type statistics. *Journal of Statistical Computation and Simulation*, 11(3-4):261–272.

[79] Ranganathan, A. (2010). Pliss: Detecting and labeling places using online change-point detection. *Robotics: Science and Systems VI*.

[80] Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology*, 46(6):900–915.

[81] Renaud, E., Large, A., Werey, C., and Barbier, R. (2013). Les échelles des systèmes d'information et d'indicateurs de l'eau destinée à la consommation humaine : intérêts et rôles d'un échelon départemental ou supra-local. In *Congrès de l'ASTEE*, page 20 p., Nantes, France. Congrès de l'ASTEE, Nantes, FRA, 04-/06/2013 - 07/06/2013.

[82] Rigaill, G., Lebarbier, E., and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and computing*, 22(4):917–929.

[83] Roocks, P. (2016). Computing Pareto Frontiers and Database Preferences with the rPref Package. *The R Journal*, 8(2):393–404.

[84] Sadani, S., Abdollahnezhad, K., Teimouri, M., and Ranjbar, V. (2019). A new estimator for weibull distribution parameters: Comprehensive comparative study for weibull distribution. *arXiv preprint arXiv:1902.05658*.

[85] Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., and Kaufman, J. D. (2011). Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, 45(36):6593–6606.

[86] Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

[87] Shi, X., Beaulieu, C., Killick, R., and Lund, R. (2022). Changepoint detection: An analysis of the central england temperature series. *Journal of Climate*, pages 1 – 46.

[88] Shipley, B. (2016). *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press.

[89] Shoari, N. and Dubé, J.-S. (2018). Toward improved analysis of concentration data: embracing nondetects. *Environmental toxicology and chemistry*, 37(3):643–656.

[90] Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98.

[91] Tartakovsky, A. G. and Moustakides, G. V. (2010). State-of-the-art in bayesian change-point detection. *Sequential Analysis*, 29(2):125–145.

[92] Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

[93] van Os, B. and Meulman, J. (2004). Improving dynamic programming strategies for partitioning. *Journal of Classification*, 21(2):207–230.

[94] Verbeke, J. and Cools, R. (1995). The newton-raphson method. *International Journal of Mathematical Education in Science and Technology*, 26(2):177–193.

[95] Vert, J.-p. and Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group lars. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

[96] Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

[97] Wang, Y., Wang, Z., and Zi, X. (2019). Rank-based multiple change-point detection. *Communications in Statistics - Theory and Methods*, 49(14):3438–3454.

[98] Wasson, J., Chandesris, A., Pella, H., and Blanc, L. (2002). Définition des hydro-écorégions françaises métropolitaines. Approche régionale de la typologie des eaux courantes et éléments pour la définition des peuplements de référence d'invertébrés. Technical report, irstea.

[99] Yang, T. Y. and Kuo, L. (2001). Bayesian binary segmentation procedure for a poisson process with multiple changepoints. *Journal of Computational and Graphical Statistics*, 10(4):772–785.

[100] Yao, Y.-C. (1988). Estimating the number of change-points via schwarz' criterion. *Statistics and Probability Letters*, 6(3):181–189.

[101] Zhang, N. R. and Siegmund, D. O. (2006). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

[102] Zhao, X. and Chu, P.-S. (2010a). Bayesian changepoint analysis for extreme events (typhoons, heavy rainfall, and heat waves): An rjmcmc approach. *Journal of Climate*, 23(5):1034–1046.

[103] Zhao, X. and Chu, P.-S. (2010b). Bayesian changepoint analysis for extreme events (typhoons, heavy rainfall, and heat waves): An RJMCMC approach. *Journal of Climate*, 23(5):1034–1046.

[104] Zheng, K., Tan, L., Sun, Y., Wu, Y., Duan, Z., Xu, Y., and Gao, C. (2021). Impacts of climate change and anthropogenic activities on vegetation change: Evidence from typical areas in china. *Ecological Indicators*, 126:107648.

[105] Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3).

Appendices

A. Proofs of Chapter 4

This section discuss if the chosen estimators prensented in Chapter A have satisfying proprieties in the context of left censored data. This section consists in verifying all the conditions and hypothesis needed to ensure the convergence of our chosen estimators. First, we check if $\hat{\lambda}$ estimated with the Newton-Raphson procedure converges to λ^*. Then, we will follow the same demonstration than [53] to proove that the left censorship does not hinder the convergence of the breakpoint estimators $\hat{K}$ and $\hat{t}$. Lastly, we will check if the necessary conditions to use the PELT algorithm are verified.

A.1. On the convergence of $\hat{\lambda}$

We managed to find a computational way to treat samples where all measurements are censored. In that case, the maximum is reached when λ tends to infinity (there is no convergence) and the cost function tends to 0. In our algorithm, the value of $\hat{\lambda}$ is upper bounded by a $\lambda_{max} = 10^6$. We are interested in the other case where there are some non censored observations. The scale estimator $\hat{\lambda}$ is estimated using Newton-Raphson. The convergence hypothesis stated in *Theorem 1* of [94] are hard to verify given the formula of the function to be maximized. Since we are looking for the maximum of Equation 4.7, it implies to find the zero of Equation 4.15. We can rewrite this equation :

$$\frac{\partial \sum_{i=t_k+1}^{t_{k+1}} \ln f(y;\lambda,\sigma)}{\partial \lambda} = n_a \frac{a^\sigma \sigma(\lambda)^{\sigma-1} e^{-(\lambda a)^\sigma}}{1-e^{-(\lambda a)^\sigma}} + (n_{seg}-n_a)\frac{\sigma}{\lambda} - \sigma(\lambda)^{\sigma-1}\sum_{i \notin \mathcal{N}_a} y_i^\sigma, \tag{A.1}$$

with n_{seg} being the number of observations in segment $t_k+1 : t_{k+1}$ and n_a its number of censored observations. We will denote A.1 by $\partial\mathcal{L}(\lambda; y, \sigma)$ to simplify the notations. We can then compute the second derivative that can be written :

$$\begin{aligned}\frac{\partial^2 \sum_{i=t_k+1}^{t_{k+1}} \ln f(y;\lambda,\sigma)}{\partial \lambda^2} = n_a \frac{a^\sigma \sigma(\sigma-1)\lambda^{\sigma-2} e^{-(a\lambda)^\sigma}}{1-e^{-(a\lambda)^\sigma}} - n_a \frac{(a\sigma\lambda^{\sigma-1})^2 e^{-(a\lambda)^\sigma}}{(1-e^{-(a\lambda)^\sigma})^2} \\ - (n_{seg}-n_a)\frac{\sigma}{\lambda^2} - \sigma(\sigma-1)\lambda^{\sigma-2}\sum_{i \notin \mathcal{N}_a} y_i^\sigma,\end{aligned} \tag{A.2}$$

that we will denote $\partial^2\mathcal{L}(\lambda; y, \sigma)$.
The iteration function ϕ described in [94] can be written as:

$$\phi(x_n) = x_n - \frac{\partial\mathcal{L}(\lambda; y, \sigma)}{\partial^2\mathcal{L}(\lambda; y, \sigma)} \tag{A.3}$$

We need to show that ϕ is a contraction to prove the convergence of the Newton-Raphson method. We can see that it is not obvious at first glance since the variations of these functions

depend on various factors such as the value of σ, the number of censored observations n_a in the sample. For instance, we considered in all simulations of Chapter 4 that $\sigma \leq 1$ which changes the sign of all the terms depending on $\sigma - 1$ in A.1 and A.2. All we can safely say is that those functions are continuous in λ. We represented for different configurations of λ, censorship rate and σ the variations in λ of the function ϕ. We show here a simple scenario where the real sample scale parameter $\lambda^\star$ takes the values 2 and 0.01. We explore the variation of around the true value of $\lambda^\star$ on the interval $[\lambda^\star - \frac{\lambda^\star}{5}, \lambda^\star + \frac{\lambda^\star}{5}]$. All results are presented in Figure A.1. We can suppose that ϕ is a k-Lipschitz function for $k \leq 1$ however the influence of the censorship rate is very important. The more censored observations are present in the sample the more the interval $[\lambda^\star - \frac{\lambda^\star}{5}, \lambda^\star + \frac{\lambda^\star}{5}]$ is not the window where the ϕ is $k-$Lipschitz. This stresses also the importance of the initialisation value of λ in the Newton-Raphson procedure.

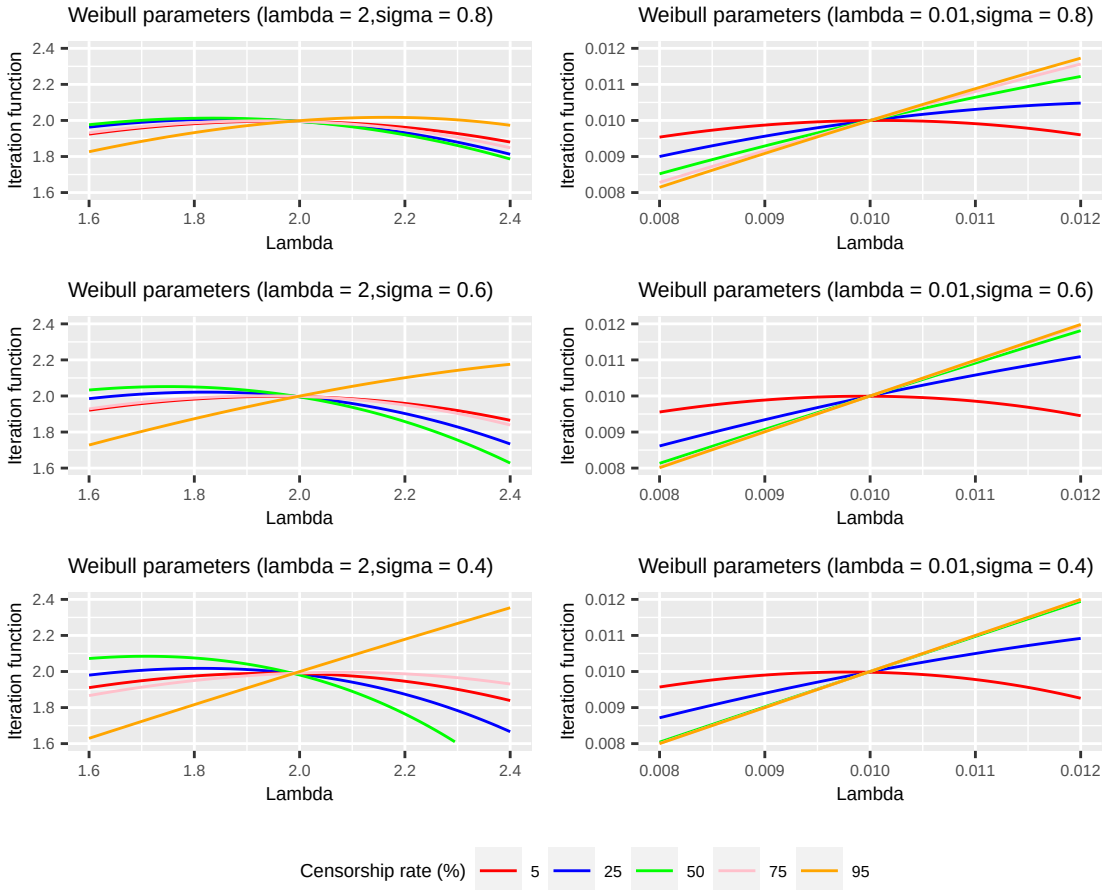

Figure A.1: Iteration function plot. The values of y_i used was drawned from a 1000000 size samples of Weibull realisations with parameters $(\lambda^\star, \sigma)$

We also provide in Figure A.2 the variations of the second derivative $\partial^2 \mathcal{L}(\lambda; y, \sigma)$. We can assume that it is strictly positive as stated in Chapter 4 ensuring the existence of a minimum in the neighbourhood of $\lambda^\star$.

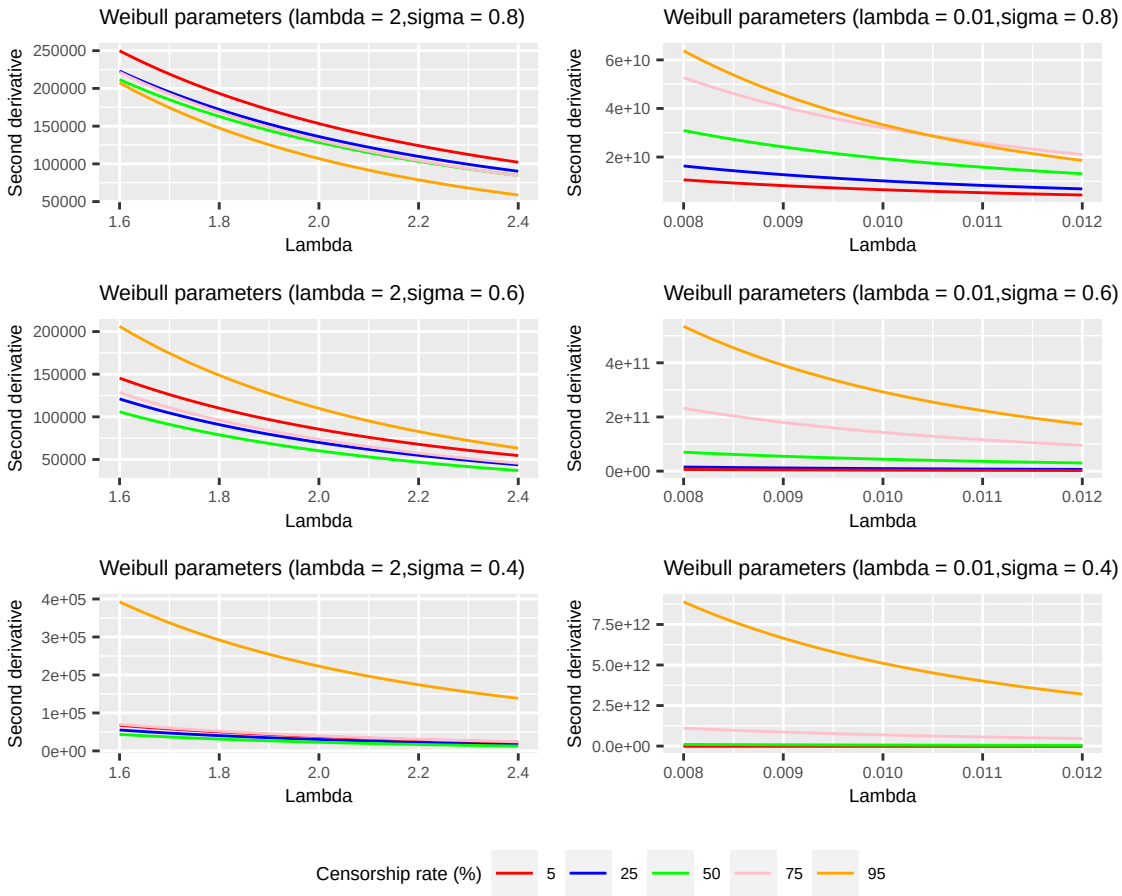

Figure A.2: $\partial^2\mathcal{L}(\lambda; y, \sigma)$ plot. The values of y_i used was drawned from a 1000000 size samples of Weibull realisations with parameters $(\lambda^\star, \sigma)$.

A.2. Convergence of $\hat{K}$ and $\hat{t}$

We provide here the proof of proposition 4.9. We base our demonstration entirely on the approach developped in [53]. It is assumed that (H1)-(H3) are true. It remains for us to introduce another condition. But first let's denote $\eta_i = \ln f(Y_i, \lambda_k, \sigma) - \mathbb{E}[\ln f(Y_i, \lambda_k, \sigma)]$ for i belonging to the k-th segment and associated to the parameters λ_k and σ. We have the following proposition:

Proposition A.2.1. *There exists $C < \infty$ such that for any $t \geq 0$ and any $s > 0$,*

$$\mathbb{E}[\sum_{i=t+1}^{t+s} \eta_i]^2 \leq Cs^h, \tag{A.4}$$

for some $1 \leq h \leq 2$.

This condition corresponds to the condition C0(h) of [53] and is quite common; it is also be found as an assumption in [40]. However, it is [53] that gives us the indication that this

condition is indeed verified in our case. It explains that in the application framework, if the base signal of $(Y_1, ..., Y_n)$ is generated by independent variables, then the variable η_i defined in A.4 is also a sequence of random variables and the proposition is verified even verified for $h = 1$. Then from Theorem 2.2 of [53], we have the consistency of the estimator:

$$(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\lambda}}) \xrightarrow[n\to\infty]{\mathbb{P}^\star} (\boldsymbol{\tau}^\star, \boldsymbol{\lambda}^\star)$$

A.3. Verifying PELT assumptions

Some necessary conditions must be met before using the PELT algorithm. It can be found in Theorem 3.1 of [52] and can be stated as follow:

Proposition A.3.1. *We assume that when introducing a changepoint into a sequence of observations the cost, $\mathcal{C}$, of the sequence reduces. More formally, we assume there exists a constant K such that for all $t < s < T$,*

$$W(y_{t:s}) + W(y_{s:T}) + K \leq W(y_{t:T}) \tag{A.5}$$

Then if

$$F(t) + W(y_{t:s}) + K \geq F(s) \tag{A.6}$$

holds, at a future time $T > s$, t can never be the optimal last change point prior to T.

Proof: The equation A.5 is always verified with working with additive criterion such as the log likelihood. We can see that in the case of our cost function:

$$W(y_{t:s}, \hat{\lambda}_{t:s}) + W(y_{s:T}, \hat{\lambda}_{s:T}) + K \leq W(y_{t:T}, \hat{\lambda}_{t:T})$$

It is a direct consequence of using the maximum likelihood estimator. Suppose now that A.6 is true. Adding $W(y_{s:T}, \hat{\lambda}_{s:T})$ on both sides of the inequation gives :

$$\begin{aligned} F(t) + W(y_{t:s}, \hat{\lambda}_{t:s}) + W(y_{s:T}, \hat{\lambda}_{s:T}) + K &\geq F(s) + W(y_{s:T}, \hat{\lambda}_{s:T}) \\ \implies F(t) + W(y_{t:T}, \hat{\lambda}_{t:T}) &\geq F(s) + W(y_{s:T}, \hat{\lambda}_{s:T}), \end{aligned}$$

We can conclude that the segmentation with the smallest cost is the one with s as the last change-point. So t cannot be the last change-point prior to T.

B. Complement of Chapter 5

B.1. Simulation on the convergence of σ

The experimental protocol is the following. We simulated $N = 100$ samples of size $n = 320$ of left censored Weibull realisations. 3 change points are present in the samples at position 80, 160 and 240. The associated parameters for each segment are $(1, 1/100, 1/100, 1)$. Four scenarios are proposed where the σ^* and the censoring rate α varies. We test all configuration possible for $\sigma^* = (0.4, 0.8)$ and $\alpha = (25\%, 75\%)$. All the results are presented in Figures B.1 and B.2.

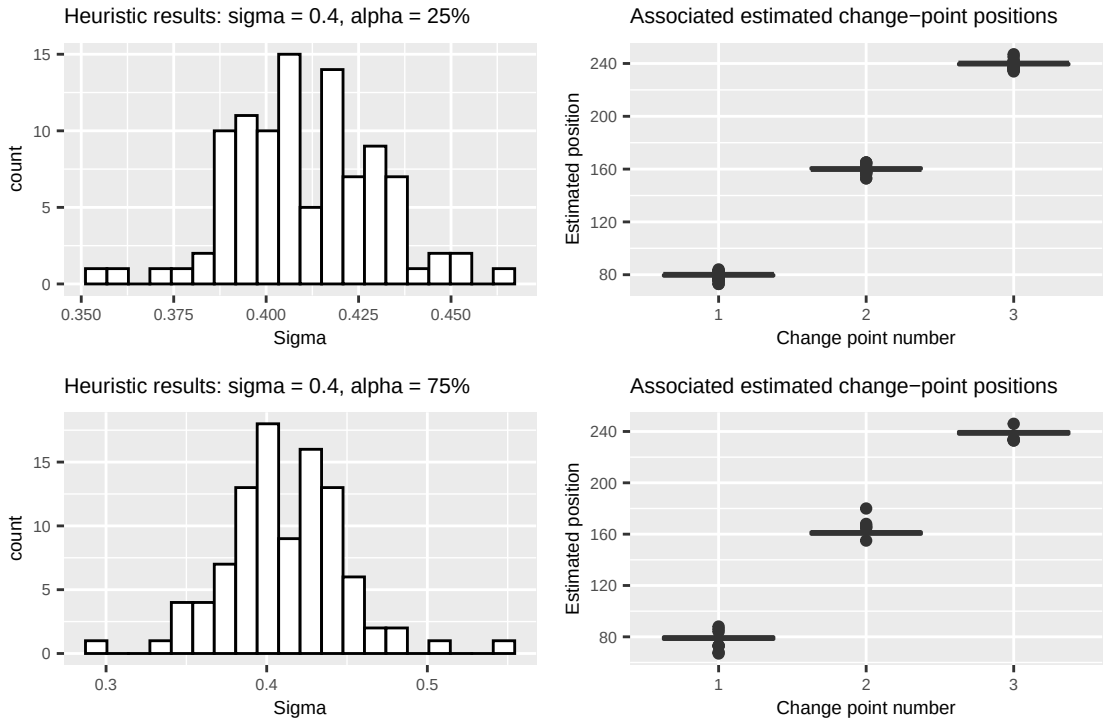

Figure B.1: Scenarios with $\sigma = 0.4$.

For each of these samples, we use the heuristic proposed in 5.2.1. The penalty grid was defined by $Q = 5$ values set to $[\beta_0 = \frac{\ln n}{10}, \ldots, \beta_q, \ldots, \beta_Q = 5 \ln n]$ with the β_q being equidistant. We allowed this important range in the penalties to ensure enough distinct points for the elbow heuristic. The only PELT parameter that wasn't set as in Chapter 4 was the minimal segment size. It was set to 25. The choice was motivated by the computational time of the simulations, even though we have seen that the estimation of the detection capacity of our method is lower with this minimal segment size. In the procedure, the first iteration $\widehat{\sigma}_0$ is computed with the `fitdistr` R package [23]. In the second step of the heuristic, minimizing 5.10 with respect to σ

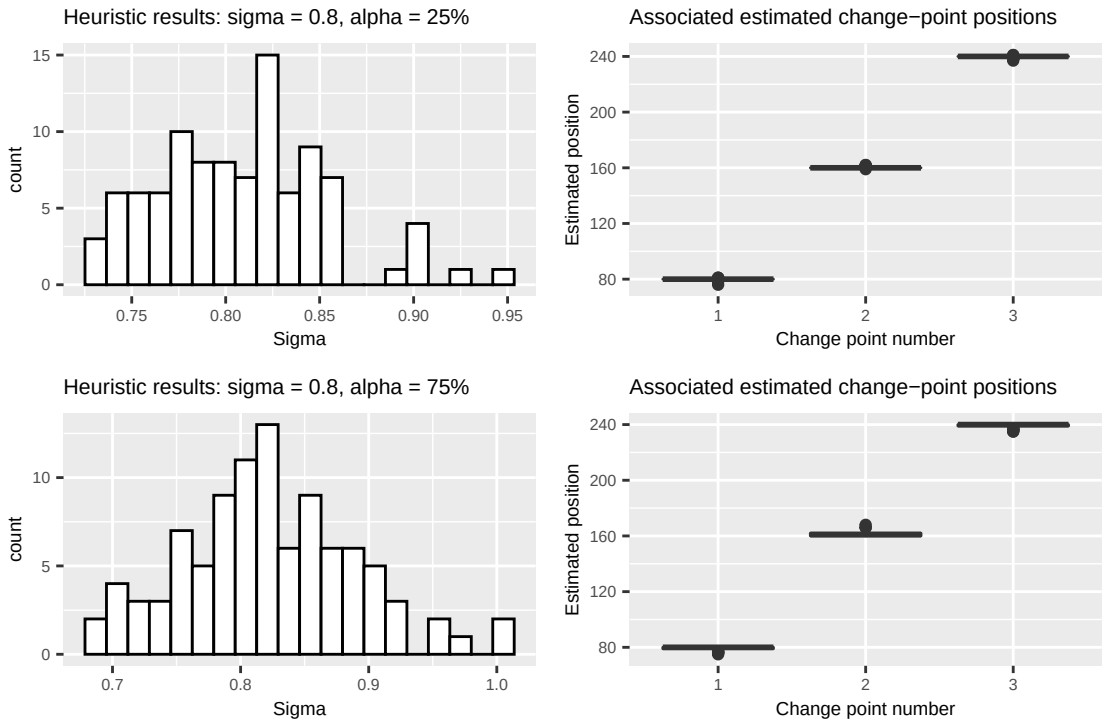

Figure B.2: Scenarios with $\sigma = 0.8$.

is done using [13] which allows for a box constraint for the parameter value σ (lower and upper bounds). This interval was set to $[0, 1]$. This explains that the estimated values are all inferior to 1 in Figure B.2. We defined a stopping criterion to the heuristic depending on the $\widehat{\sigma}$ values that consists in stopping when the upgrade of the new $\widehat{\sigma}$ is not superior to $10^{-}3$.

B.2. Clustering algorithms

We keep the same notations than in 5.2.1. We introduce the following new notations :

- C_m^p the m-th cluster located in component p.
- M_p the number of clusters in component $\mathcal{K}_p$.
- $Q(\mathcal{K}_p, C_m^p) = \dfrac{1}{|C_m^p|} \displaystyle\sum_{v_i, v_j \in C_m^p} d_{ij}^2$ the inertia of cluster C_m^p.
- $R_p(M_p) = \min\limits_{(C_m^p)_{m=1}^{M_p}} \displaystyle\sum_{m=1}^{M_p} Q(\mathcal{K}_p, C_m^p)$ the best partition (in the sense of minimal inertia) of component p into M_p clusters.

- $S(l,m) = \min_{(M_p)_{p=1}^{l} \text{ such that } \sum_{p=1}^{l} M_p = m} \sum_{p=1}^{l} R_p(M_p)$ which is the best partition of the l first components into a total number of m clusters.

$R_p(m)$ can be computed with Ward hierarchical clustering technique. In the case of this work we used the `R` package `hclust`. With these notations, we can write the two developed methods as follows:

Algorithm 5 Clustering with greedy method:

input : the station graph $G = (V, E)$, the known partition into non connex components $(\mathcal{K}_\infty, \ldots, \mathcal{K}_P)$, a total number of clusters M

initialisation : Compute $R_p(1)$ for all $p \in [1, \ldots, P]$ using `hclust`, set $M_{opt} = (1, \ldots, 1)$ vector of size P

for $m = 1$ to $M - P$ **do**
  $score \leftarrow (0, \ldots, 0)$ vector of size P
  for $p = 1$ to P **do**
    $M_{opt}(p) \leftarrow M_{opt}(p) + 1$
    $score(p) \leftarrow \sum_{p=1}^{P} R_p(M_{opt}(p))$
    $M_{opt}(p) \leftarrow M_{opt}(p) - 1$
  end for
  $pos \leftarrow which.min(score)$
  $M_{opt}(pos) \leftarrow M_{opt}(pos) + 1$
end for
for $p = 1$ to P **do**
  built the optimal partition of $\mathcal{K}_p$ with $M_{opt}(p)$ clusters using `hclust`.
end for

Algorithm 6 Clustering by dynamic programming:

input : the station graph $G = (V, E)$, the known partition into non connex components ($\mathcal{K}_\infty$, a total number of clusters M

for $p = 1$ to P **do** :
  Use `hclust` to compute $R_p(m)$ for all $m \in \{1, \ldots, M - P + 1\}$
end for
for $m = 1$ to $M - P + 1$ **do** :
  $S(1, m) \leftarrow R_1(m)$
end for
for $l = 2$ to P **do** :
  for $m = l$ to M **do** :
    $W(l, m) \leftarrow 1$
    $S(l, m) \leftarrow S(l - 1, m - 1) + R_l(1)$
    for $u = 1$ to $m - l + 1$ **do**
      if $S(l - 1, m - u) + R_l(u) < S(l, m)$ **then**
        $W(l, m) \leftarrow u$
        $S(l, m) \leftarrow S(l - 1, m - u) + R_l(u)$
      end if
    end for
  end for
end for
$M_{opt} \leftarrow (\mathrm{NA}, \ldots, \mathrm{NA})$
$P_{opt}(P) \leftarrow W(P, M)$
$left \leftarrow M - W(P, M)$
for $p = P - 1$ to 1 **do**
  $P_{opt}(p) \leftarrow W(p, left)$
  $left \leftarrow left - W(p, left)$
end for
for $p = 1$ to P **do**
  built the optimal partition of $\mathcal{K}_p$ with $P_{opt}(p)$ clusters using `hclust`
end for

B.3. Modified empirical Wasserstein distance

The Wasserstein distance was chosen over the Kolmogorov-Smirnov or the Jensen-Shannon metric. It has the advantage of integrating in the distance calculation both the differences between the probabilities of observing different values but also the distances between those values. This is a critical point which is illustrated on a simple simulated example provided by Figure B.3. We show here three monitoring stations that have quite different behaviors. Those different behaviors are obvious both on in the temporal representation and in the histograms. However, the Kolmogorov-Smirnov distance between stations 1 and 3 is equal to the Kolmogorov-Smirnov distance between stations 1 and 2. This distance cannot capture the fact that station 2 recorded higher concentration values than station 3. On the contrary, the Wasserstein distance between stations 1 and 3 is smaller than the Wasserstein distance between stations 1 and 2.
Computing information theoretic distances/dissimilarities such as the Jensen-Shannon divergence requires estimating densities for the distributions observed at the stations. As noted earlier, few concentration records (and even fewer quantified ones) are available at the level of a station and within a time period. Therefore, density estimations based on such a small number of observations are unreliable.
The empirical 1-d Wasserstein distance used in our work is slightly adapted for left censored values. Given two samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_m)$ of sizes n and m with respective empirical c.d.f. F_n and G_m, the 1-d empirical distance writes:

$$W_1(F_n, G_n) = \int_{\mathbb{R}} |F_n(x) - G_m(x)| \mathrm{d}x$$

In the case of left censored observations, the empirical c.d.f. the first non zero value is the censoring threshold. If we use the classical empirical c.d.f., it does not take into account that the potential real values of censored samples is potentially lower than this threshold. In particular, if both samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are fully censored at respective thresholds a_1 and a_2, the Wasserstein distance equals $|a_1 - a_2|$. We would like this quantity to be the smallest possible since none of the samples has any quantified values. Since the samples size for a single station is usually very small, a reasonable assumption is to suppose that the real values under the censoring threshold are uniformly distributed. Figure B.4 illustrates the changes it implies on the empirical c.d.f.. In the previous example of $\boldsymbol{x}$ and $\boldsymbol{y}$, the adapted empirical distance gives $|a_1 - a_2|/2$.

B.4. Supplementary Figures

B.4.1. Regional map of crops

The regional map of crops provided in Figure B.5 have been produced using data from the *registre parcellaire graphique* produced by the IGN [45].

B.4.2. Prosulfocarb sales

Prosulfocarb sales figures used to build Figure B.6 are made available by the *Système d'information sur l'eau* [75].

B.4.3. All elbow methods figures

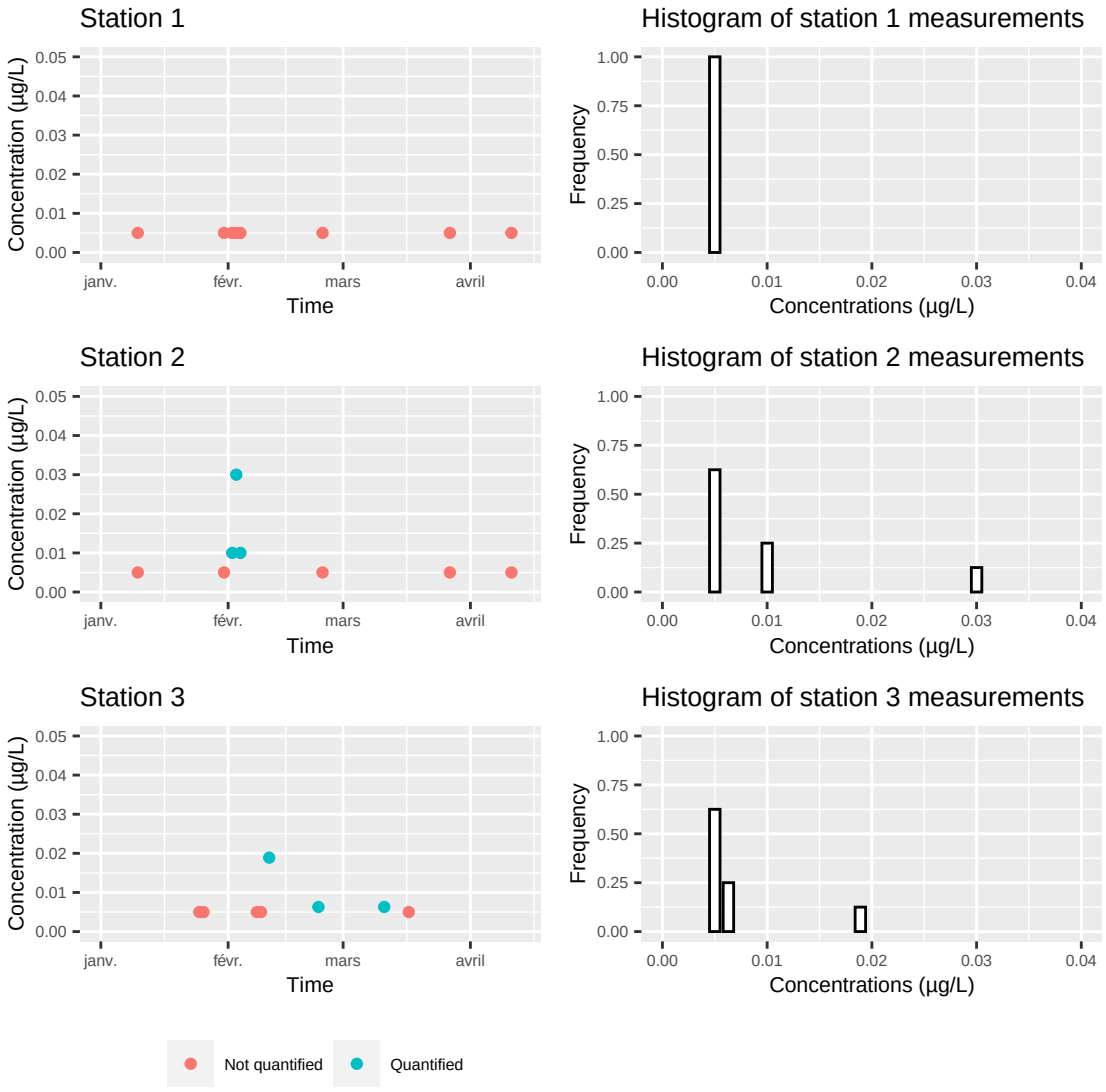

Figure B.3: Example of three stations data. The data were simulated.

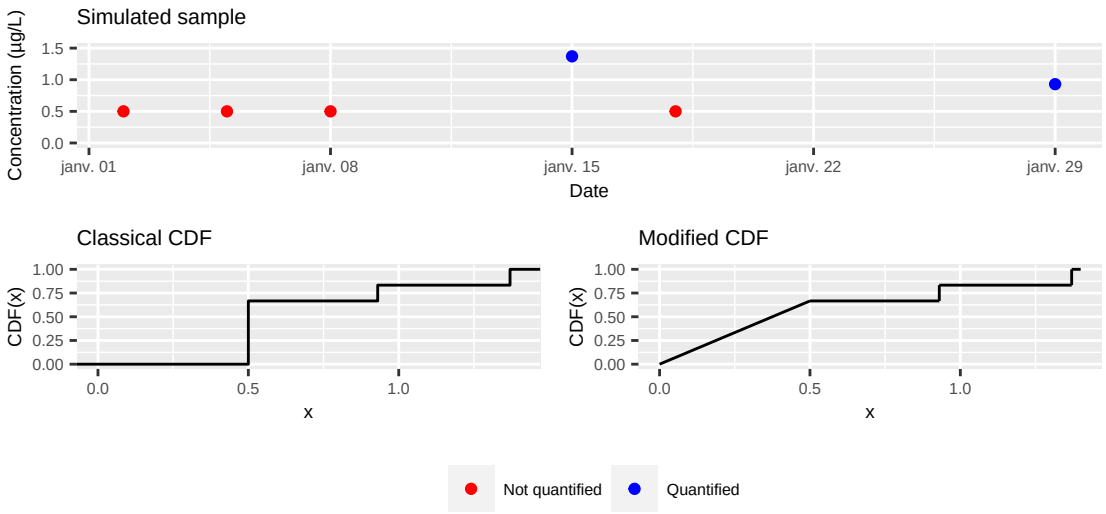

Figure B.4: Example of modified c.d.f. for the Wasserstein distance.

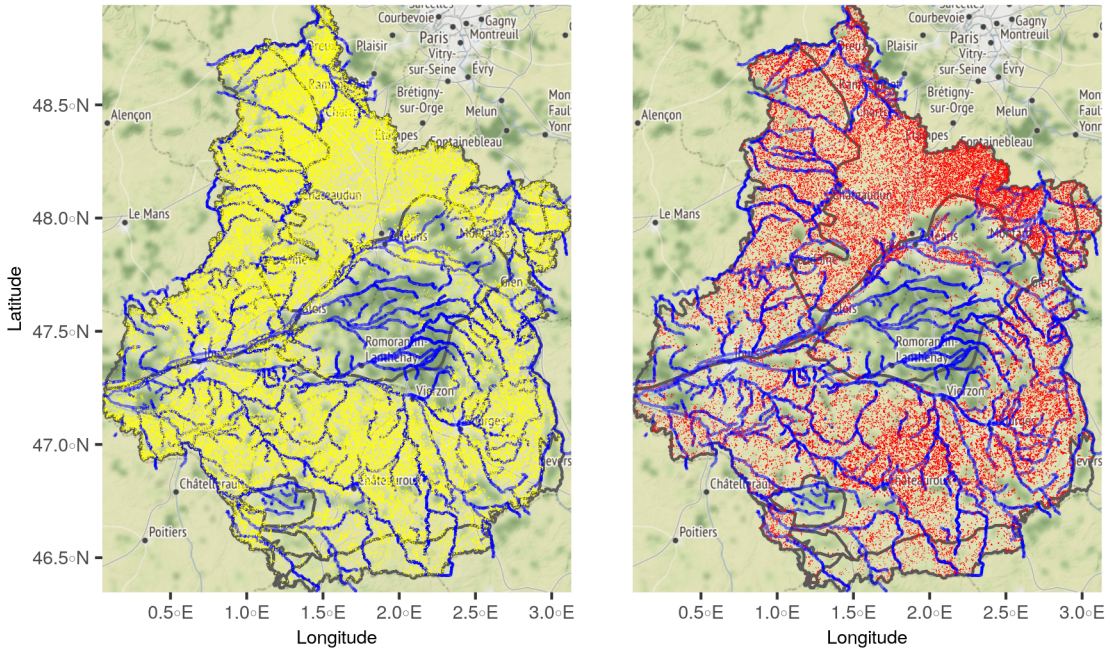

Figure B.5: Wheat (in yellow) and barley (in red) crops location in Centre-Val de Loire

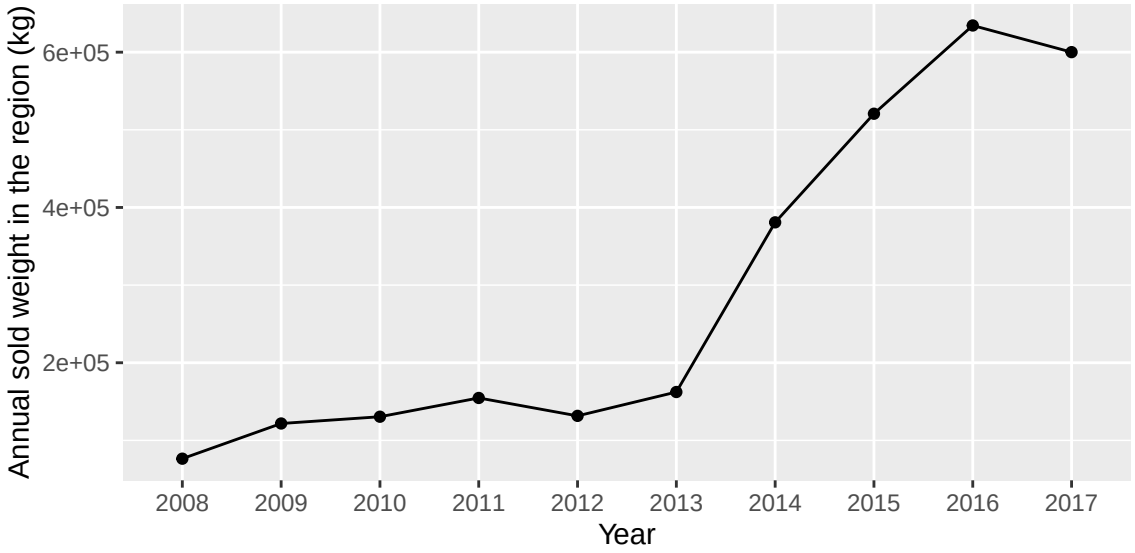

Figure B.6: Prosulfocarb sales between 2008 and 2017 in the Centre-Val de Loire region

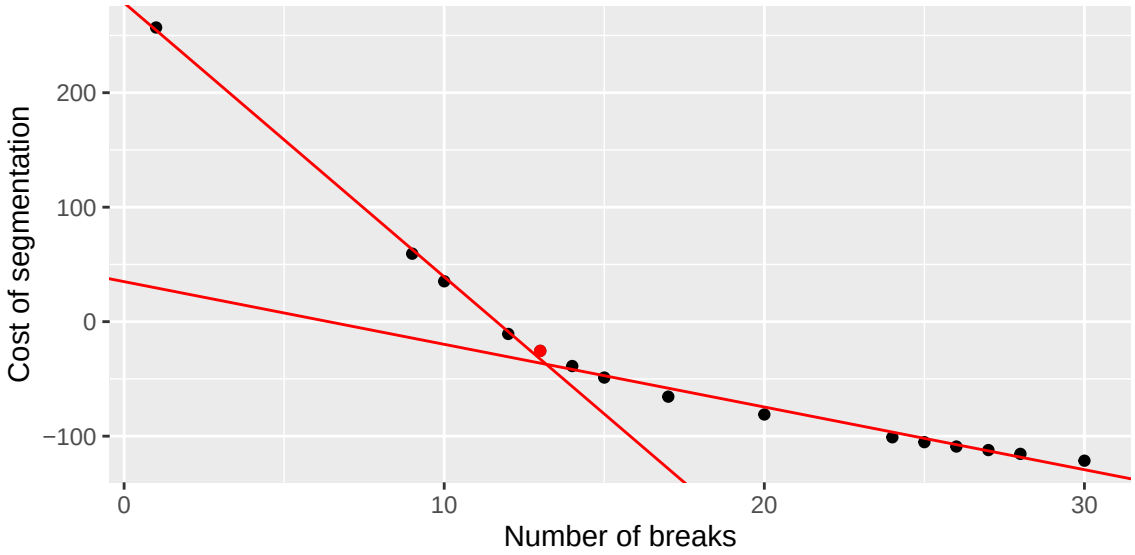

Figure B.7: Elbow method selecting the optimal segmentation of the full signal $\overline{\mathcal{D}}$.

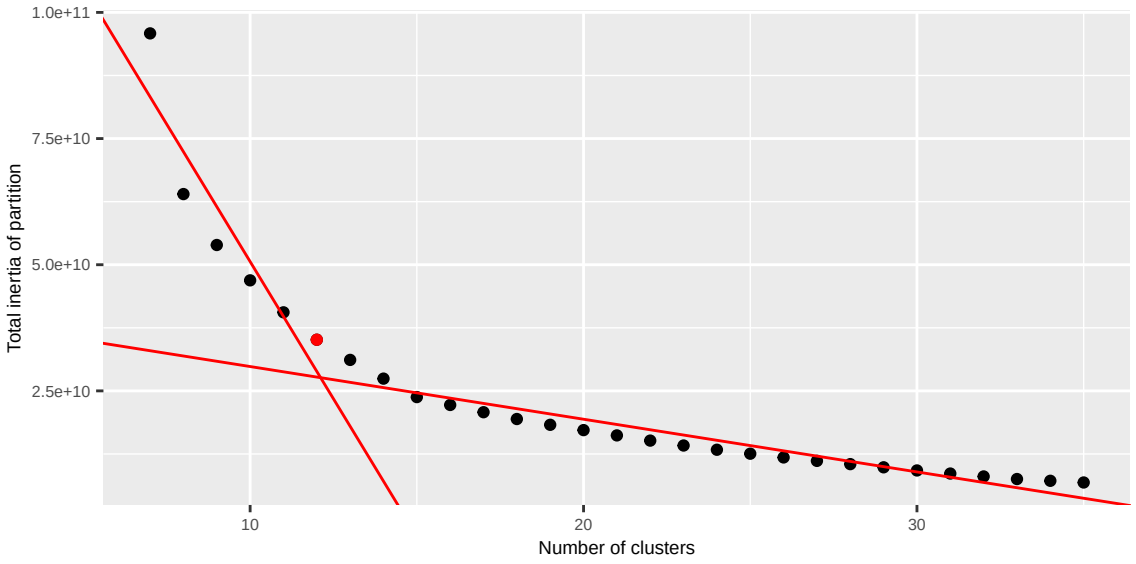

Figure B.8: Elbow method for the spatial clustering.