Pesticides concentration monitoring from various heteregenous sources of information

Clément Laroche

Pesticides concentration monitoring from various heteregenous sources of information

A statistical approach integrated into an interactive tool

Author: Clément LAROCHE

June 1, 2023

Keywords:
Pesticides, environmental study, spatio-temporal heterogeneity, change point detection, spatial clustering, anomaly detection

Conducted under the supervision of:

M. Fabrice ROSSI	-	Université Paris-Dauphine, PSL University
Mme. Madalina OLTEANU	-	Université Paris-Dauphine, PSL University

Jury

Mme. Rebecca Killick	Lancaster University	Rapportrice
M. Laurent Oudre	ENS Paris Saclay	Rapporteur
Mme. Cécile Hardouin	Université Paris Nanterre	Examinatice
M. Alain Celisse	Université Paris 1 Panthéon Sorbonne	Examinateur
M. Josselin Rety	Anses	Membre invité

Préambule

Ce rapport a fait l'objet d'un financement de l'ANSES au titre du le dispositif de phytopharmacovigilance (https://www.anses.fr/fr/content/la-phytopharmacovigilance) et dans le cadre d'une convention de Recherche et de Développement (2018-CRD-15_PPV18). Les conclusions exprimées dans ce rapport n'engagent que leurs auteurs. Le contenu de ce rapport ne fait pas l'objet d'une validation de l'Anses et ne reflète pas nécessairement son avis définitif. La responsabilité de l'Anses ne peut être engagée en cas d'informations incomplètes ou erronées. Toute utilisation ou modification de ce rapport par des tiers, sous quelque forme que ce soit, est faite sous leur seule et entière responsabilité, sans que celle de l'Anses ne puisse être recherchée.

Maitre d'œuvre : Université Paris 1 Panthéon-Sorbonne
Contact responsable de l'étude : Fabrice Rossi (fabrice.rossi@dauphine.psl.eu)

Maitre d'ouvrage (ANSES) : DER/UPPV

Nature de la ressource : *document*
Droits d'accès : *restreint*
Niveau de lecture : *expert*

Version	Parution	Modifications
Finale	*13/10/22*	-

Résumé

La mission de phyto-pharmacovigilance consiste à établir la surveillance de concentrations des produits phyto-pharmaceutiques dans des milieux environnementaux d'intérêt. Établir une telle surveillance n'est pas chose facile étant donné l'importante accumulation de données de natures différentes qui pourraient aider à une meilleure compréhension de la diffusion des substances surveillées. Les données de concentrations sont relevées par des stations de mesure réparties sur l'ensemble du territoire. Ce sont donc des données spatio-temporelles. De plus, les données relevées présentent plusieurs caractéristiques qui compliquent leur analyse comme de la censure, de l'hétérogénéité spatio-temporelle ou encore une forme particulière dans leur distribution empirique. On cherche donc à développer des méthodes adaptées à ces caractéristiques pouvant extraire des informations spatiales et temporelles anormales à fournir à une équipe d'expert de la phyto-pharmacovigilance.
Deux contributions originales sont développées dans ce manuscrit de thèse. La première est une méthodologie en trois étapes dont la finalité est de détecter des clusters anormaux lors de périodes temporelles précises. Les périodes temporelles en question sont obtenues par détection de ruptures sur la série des concentrations agrégée à une certaine échelle temporelle. Une fois que l'on se place dans un segment temporel découlant de cette détection, on peut effectuer une comparaison de clusters spatiaux obtenus par clusterting hiérarchique. Cette comparaison peut être implémentée par optimisation multi-critère. On peut de cette manière détecter des clusters anormaux de manière contextuelle au segment temporel sélectionné. La méthode de détection de ruptures est spécialement adaptée aux caractéristiques des données de concentrations. Les performances de la détection de ruptures sont testées sur des données simulées. Elle est également comparée à une méthode de la littérature adaptée aux données censurées. La deuxième contribution de ce travail est une présentation interactive de l'ensemble des résultats obtenus par cette méthode sous la forme d'une application interactive `Rshiny`.

Mots-Clés : phyto-pharmacovigilance, détection de ruptures, clustering, optimisation multi-critère, données spatio-temporelles, données censurées à gauche, hétérogénéité spatio-temporelle, application `Rshiny`.

Abstract

The task of phytopharmacovigilance is to establish monitoring of concentrations of phytopharmaceuticals in relevant environmental media. Establishing such monitoring is not an easy task given the large accumulation of data of different types that could help to better understand the distribution of the monitored substances. Concentration data are collected from monitoring stations distributed throughout the area. It is therefore spatio-temporal data. Moreover, the collected data have several characteristics that complicate their analysis, such as censoring, spatial and temporal heterogeneity and a particular form of their empirical distribution. We are therefore trying to develop methods adapted to these characteristics that can extract anomalous spatial and temporal information to make it available to a team of phytopharmacovigilance experts.
Two original contributions are developed in this manuscript. The first is a three-step methodology aimed at detecting anomalous clusters in specific time periods. The time periods in question are identified by detecting change points in the aggregate concentration series on a given time scale. In a time segment resulting from this detection, a comparison of the spatial clusters obtained by hierarchical clustering can be performed. This comparison can be performed by multi-criteria optimisation. Clusters are highlighted as anomalous if they have a high Pareto front value. Anomaly detection is thus contextual to the selected temporal segment. The break detection method is specifically adapted to the characteristics of the concentration data. The performance of the break detection is tested on simulated data. It is also compared with a method from the literature adapted to censored data. The second contribution of this work is an interactive presentation of all results obtained with this method in the form of an interactive application `Rshiny`.

Keywords : phytopharmacovigilance, change point detection, clustering, multi-criterion analysis, spatio-temporal data, left censored data, spatio-temporal heterogeneity, `Rshiny` application.

Contents

List of Tables

List of Algorithms

List of Figures

1. Introduction

This thesis is the main result of a research and development agreement between the French Agency for Food, Environmental and Occupational Health and Safety (ANSES[1]) and the laboratory SAMM[2] of the University Paris 1 Pantheon-Sorbonne. One of the main tasks of ANSES is to monitor environmental data and provide tools which ensure the protection of the public and the environment. In particular, national health authorities are very interested in monitoring and quantifying the concentration of various pollutants in specific environmental areas, as nowadays significant adverse health effects are well documented (Khopkar (2007); Marchant et al. (2018); Nougadère et al. (2014) for example). Within the Agency, the Phytopharmacovigilance Unit (UPPV) is entrusted with this task. Indeed, the role of this unit is to establish a list of pollutants that must be monitored according to their potential harmful effects on the environment and health, and ultimately to issue recommendations for their use, in terms of dosage, duration and precautions of use. To achieve this goal, UPPV leads the national network of phytopharmacovigilance, which monitors environmental concentrations of phytopharmaceutical products.

Over the years, considerable efforts have been made at national level to build up public databases storing relevant information such as phytopharmaceutical product concentrations in surface waters (Office français de la biodiversité, n.d.) and hydrographic characteristics of French rivers (Institut National de l'Information Géographique et Forestière, n.d.). As a result, there is now a considerable amount of data available, for which basic visualization and descriptive statistics are currently performed (see chapter 2 for an example). UPPV now aims to make better use of these databases, which have not yet been fully exploited, by defining new algorithms and statistical methods to achieve better pollution monitoring. The objective is to help experts summarise and analyse information from field measurements, evaluate these measurements, detect possible anomalies and issue warnings if necessary. An example of such a tool could be, for example, a monitoring system designed to detect patterns such as fluctuations, trends or changes in relevant series of environmental measurements over time and space (Manly, 2008).

The work accomplished during this thesis takes place in this context, all the discussions and developments presented in this manuscript have been carried out in close cooperation with the experts of UPPV in order to address the objective mentioned above, more precisely the development of a method to detect changes and anomalous episodes in environmental measurements over time and space, as well as their numerical implementation.

Environmental concentrations of phytopharmaceutical products are collected through a network of measuring stations distributed over the entire French territory. Modelling these data is a complex issue for several reasons. First of all, the monitoring involves the use of sensors at

[1] Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail

[2] Statistique, Analyse et Modélisation Multidisciplinaire

different sites that take samples at different times, so that the data collected contain spatio-temporal information. Other difficulties are related to the characteristics of the data collected, which may in part come from limitations in the measurement process. For instance:

- Pollutant concentrations are measured with sensors, that usually have detection and quantification limits: The corresponding data are then left-censored.

- The data distribution in itself is usually skewed to the right, with long tails indicating the possible occurrence of high concentrations.

- In many situations, data are collected irregularly due to measurement practises (measurements require human intervention).

- Pollution is monitored at different locations, and different sensors may be used at each location, resulting in significant spatial heterogeneity.

This thesis proposes a new approach for dealing with spatio-temporal data that exhibit these characteristics. The basic ideas underlying this work are the following. First of all, due to the characteristics of the sampling and the variability of the environmental conditions, the concentration measurements are considered independent and drawn from distributions that are locally stationary in time. Although this assumption may be considered strong or simplistic, it gave satisfactory results in practice. From a purely temporal point of view, the analysis of these data is therefore a temporal segmentation problem in order to identify segments where the distribution of the data is stationary and homogeneous. Furthermore, the spatial structuring of the measuring stations and their positioning on river networks leads to the second idea, which is the existence of spatial zones in which a homogeneous temporal behaviour is expected. These zones can be determined by clustering approaches. Finally, the third idea is to combine the results of the temporal and spatial segmentation to systematically highlight areas with anomalous behaviour over a given time period.
The manuscript is organized as follows:

- **Chapter 2** is a presentation of the French national agency in charge of monitoring pollution data. A short description of the agency and of its missions is given with an extensive description of the available data derived from different sources of information that are useful for the analysis of environmental pollution. The last section presents a first analysis of the data sets used throughout the thesis based on descriptive statistics and visualization techniques that allows to extract some information.

- **Chapter 3** provides a presentation of specific change-point detection methods. Extensive reviews on change-point detection methods are already available in the literature, hence we focus in this chapter on features that we believe could be useful for applications to environmental data. In particular, we cover the case where the number of changes is unknown, and review search methods (or heuristics) that give an optimal solution to that problem. The last section discusses applications of change-point methods in environmental data that were previously published in the literature.

- **Chapter 4** presents a first methodological contributions of the thesis. We consider general parametric, left censored distributions (characteristic of environmental data) and study the impact of censoring on the change-point detection. Then we discuss different optimization strategies to estimate the parameters of the model. Finally, we propose an original change-point detection method, involving an iterative procedure to estimate to estimate both piecewise constant parameters and parameters that remain constant over time in the model. Simulation experiments are led to compare the proposed method with a non parametric method that is also suited for censored data.

- **Chapter 5** is the second methodological contribution of our manuscript. We propose to use the results of the change-point detection method developed in Chapter 4 on the temporal dimension of concentration data. The spatial dimension is handled using classical hierarchical clustering methods integrating the actual environmental properties (such as watercourses, winds etc...). We manage to extract useful information from the resulting geographical areas and time periods by performing multi-criterion analysis. This last step results in the detection of anomalous clusters. Our contribution results in the definition of a method which deals with rough data to finally provide alert about potential anomalies. We illustrate the whole procedure with a case study of prosulfocarb concentrations in the Centre-Val de Loire French region.

- **Chapter 6** presents the last contribution of this manuscript. It describes the development of an interactive application implemented in Rshiny that displays the results of the procedure defined in Chapter 5. This application is illustrated with the prosulfocarb dataset. The application allows to store the results for all the time segments detected and to explore them interactively. This application serves an operational purpose. It is specifically designed for the experts working in that area of expertise.

Contributions

- **Article:**
  Laroche, C., Olteanu, M., & Rossi, F. (2022a). Pesticide concentration monitoring: Investigating spatio-temporal patterns in left censored data. *Environmetrics*. doi: 10.1002/env.2756

- **Conferences:**
  Laroche, C., Olteanu, M., & Rossi, F. (2021). Estimation paramétrique de ruptures dans des données censurées à gauche. In *52ème journées de statistique de la société française de statistique (sfds).*
  Laroche, C., Olteanu, M., & Rossi, F. (2022b). Pesticide concentration monitoring: investigating spatio-temporal patterns in left censored data. In *15-th international conferenceon operation research.*

- `Rshiny` **application:**
  `https://github.com/Clement-Laroche/Application_ANSES`

2. Application context

Contents

One of the most recent policy responses made to address environmental issues in Europe is the European Partnership for the Assessment of Risks from Chemicals (PARC) (*Web resource: European commission*, n.d.). This partnership involves 28 different countries. This new project was positively evaluated by the European Commission in January 2022 and started on the 1st of May 2022. The main objectives of PARC are to promote European cooperation, advance research, increase knowledge of chemical risk assessment and train relevant methodological skills. Close cooperation between authorities and researchers will facilitate the translation of research results into regulatory practice.
The French Agency for Food, Environmental and Occupational Health and Safety (ANSES[1]) is not only the main French actor in this partnership, but also the coordinator of the whole partnership. This manuscript is the result of a mission funded by ANSES and aimed at supporting the Agency in its mission on French territory. In order to better understand the topics covered in this manuscript, it is necessary to have a presentation of the ANSES and its functions, as well as the means at its disposal to carry out its tasks. This chapter describes the overall context and specific issues arising from the data used. The chapter is structured as follows: in 2.1 we introduce the ANSES, in 2.2 we explain the task of pesticide monitoring, in 2.3 we focus on the characteristics of the pesticide measures, in 2.4 we describe other sources of information of interest and in 2.5 we define the intended objectives.

2.1. Presentation of the ANSES

The ANSES was created in 2010 from the merger of the French Food Safety (AFSSA) and the French Agency for Environmental and Occupational Health Safety (AFSSET). It is a public administrative establishment under the authority of the Ministries of Health, the Environment, Agriculture, Labour and Consumer Affairs.

2.1.1. Missions

The ANSES has many missions. Let us introduce the most important ones.

- **Research activities:** the ANSES contributes to the advancement of scientific knowledge on the exposure of humans, animals, plants and the environment to various hazards and risks and has the mission to improve their monitoring. The research topics focus on three areas: Animal Health and Welfare, Plant Health and Food Safety. ANSES is also involved in the development of new analytical methods and detection techniques to identify pathogens and contaminants,in the natural environment and in the production chain. This mission also includes health surveillance and alert activities. The ANSES participates in epidemiological surveillance platforms for animal health, plant health and food chain safety. Under this mandate, the ANSES is also in charge of coordinating five different surveillance systems. These systems respectively cover: toxicovigilance, food supplement surveillance, phytopharmacovigilance, which is further explored in 2.2, veterinary pharmacovigilance and occupational disease surveillance and prevention.

[1] Agence nationale de sécurité sanitaire, de l'alimentation, de l'environnement et du travail

- **Risk assessment:** the ANSES answers society's questions on potential risks arising from the consumption of food, the use of certain products or technologies, professional activities or the pollution of various environmental compartments (e.g. air, water or soil). Given the complexity of such risks, the ANSES has developed a working methodology that brings together many disciplines to provide the most comprehensive response possible. The ANSES also has the remit to assess the efficacy of veterinary medical substances and phytopharmaceuticals and the associated risks for human and animal heath and for the environment. In particular, the ANSES is responsible for issuing sales authorisations for such products and thus also has the power to withdraw products at national level (*Web resource: Anses decision site*, n.d.).
- **Public and environmental protection:** the ANSES makes recommendations to support public debates and decisions. Its activities contribute to the implementation of effective preventive and protective measures on various societal issues such as health, biodiversity and ethics. It also provides the public with access to reliable, independent and multidisciplinary scientific information. The ANSES's mission is to flexibly respond to already known or emerging, short- or long-term sanitary risks. In other words, the aim is to identify any emerging signals as quickly as possible and make recommendations, even in a context of crisis associated to scientific uncertainty. The task is then to reduce the level of uncertainty as much as possible. The recommendations are based on all available knowledge, generated by the ANSES itself or by its partners.

2.1.2. Means of action

The ANSES has the means to carry out and fund research in collaboration with the French and international scientific communities. It has 9 laboratories spread over its 16 sites on the French territory (including overseas departments). The research carried out in these facilities addresses the complex interactions between the environment, human health and animal health. The aim is to anticipate the emergence of zoonoses or animal diseases that could have an economic impact and to combat antibiotic resistance. Specifically, the main directions of this research are:

- To learn the characteristics of pathogens (such as fungi, bacteria, viruses or parasites), macro-organisms (such as insect pests or invasive plants) and chemical contaminants.
- To detect them using state-of-the-art analytical methods.
- To monitor them using powerful epidemiological methods.
- To understand the impact of animal husbandry on animal welfare and health.
- To develop useful knowledge for the development of new treatments and vaccines to prevent and control animal and plant diseases.

The 9 laboratories have been designated as reference laboratories for pathogen research under more than 100 national and international mandates (*Web resource: Anses laboratories mandates*, n.d.).

In addition to its research activities, the ANSES is also at the centre of a network of partners. It cannot afford collecting data on all the topics across its broad scope. Therefore, for each topic, it holds discussions with other organisations that can provide interesting sources of information. For example, each monitoring system coordinated by the ANSES mobilises a different set of partners. We will present the different datasets that these partners bring in Sections 2.3 and 2.4.

2.1.3. Specific organisation

Many national agencies are counterparts of ANSES. They all participate in the PARC partnership. Each of them shares data collected at national level, enabling research studies at European level. However, each agency depends on the internal policies and the organization of its own country. This leads to heterogeneity on different topics and at different levels.
For example, internal policies influence the list of products monitored, and some substances are not tested in certain countries because they are not even approved for sale. This leads to a patchwork of different substances lists at European level, as shown in Baran et al. (2022). Heterogeneity can also be observed at the national level. For example, in France it has been established that data on drinking water quality should be collected at regional level, see for example Baran et al. (2022). Each regional agency is responsible for the quality of the data it shares with ANSES. We will see the impact of this organisation on the data in Section 2.3.
The structure of the monitoring system is thus country-specific. This is especially true for the pesticide monitoring system, on which we will focus below.

2.2. Pesticides monitoring mission

Although the term risk is often confused with hazard in common usage, they do not have the same definition. A **health hazard** is the inherent ability of a substance or organism to cause adverse health effects. **Exposure** is the specific situation in which people are confronted with a health hazard. Exposure can be characterised by the following questions:

- What was the degree or intensity of exposure?
- How long and how regularly does the exposure occur?
- In what way does the exposure occur? (E.g. skin contact, ingestion, etc.)

A **health risk** occurs when one is exposed to a health hazard. It is defined as the probability of the occurrence of adverse effects on human health. It can take many forms, such as infection, poisoning or chronic disease (such as diabetes or asthma). The outcome depends on the characteristics of the exposure and the characteristics (such as age or immunity) of the animal, human or plant population studied.
The ANSES has the task of investigating and monitoring health risks caused by various factors, including chemical substances like pesticides. The pesticide monitoring mandate can be formally defined as a surveillance system that collects and evaluates data on phytopharmaceuticals

(pesticides). It can also be referred to as phytopharmacovigilance. 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. The health hazards of pesticides are well referenced by the Agency and available in the AGRITOX database (*Web resource: AGRITOX database*, n.d.). Depending on the population targeted by the monitoring, different types of exposure can be distinguished. In the following, two concrete cases are presented to illustrate the variability of exposures that can be observed in practice.
The first case concerns the monitoring of health risk for professional farmers. They are regularly exposed to the pesticides they use. The exposure is then of long duration and the likely routes of exposure to the substance would be inhalation or skin contact.
The second case is about monitoring health risk for aquatic fauna. Exposure to pesticides may result from water run-off after pesticide application and diffusion in the local river system. As the pesticide is in the same environment as the aquatic fauna, it could come into direct contact with them and cause adverse effects (acute or chronic).
It should be noted that diffusion and thus exposure both depend on external factors, independent of the intrinsic properties of the substance. One example would be the meteorological conditions during the study period. Another example would be environmental characteristics that are likely to influence the diffusion of the substance.
For example, the diffusion of a chemical product in a river system could be influenced by the composition of the river bed or the width of the river. Therefore, it would be interesting to study the variability of exposure within regions where these characteristics are fixed. The hydro-ecoregions (HER) take into account such characteristics and would provide a coherent additional analytical tool for the river system example.
The ANSES does not directly collect the data it needs to fulfil its phytoharmacovigilance mandate. It relies on its network of partners to obtain datasets of interest. 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 investigation. For phytopharmacovigilance, the following information is of interest:

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

2. 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 datasets that provide information on pesticide concentrations and use are presented in detail in Section 2.3.

2.3. Pesticide measurement data

There are two types of measurements: direct measurements, which are essentially stations that measure the exact concentration of a substance in a particular environment, and indirect measurements, which give indications of the use on a substance.

2.3.1. Direct measurement characteristics

Access to direct measurement data is often public and data can be accessed from various portals depending on the monitored environment. These include: the Naiades portal for surface water quality (*Web resource: Naiades portal*, n.d.), the Ades portal for underground water quality (*Web resource: Ades portal*, n.d.), the Geodair portal for air quality (*Web resource: Geodair portal*, n.d.).
This part examines the characteristics of measurements directly obtained from sampling stations. The following characteristics are not present systematically, but are very common when it comes to concentration measurements.

Precision limits in chemical measurements

The first specificity of concentration measurement arises from the problem of measuring a chemical substance in a sample. In applied chemistry, every measuring device is characterised by two types of limit values:

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

These two accuracy limits imply that the concentration data are left-censored.
These limits are determined by the sensors with which the sampling station is equipped. It happens that geographical areas are covered by stations that do not have the same equipment. In this case, there are several LOD and LOQ values among 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 areas larger than French administrative regions. Administrative regions can be partially included in the area covered by a water agency. 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 the measuring instruments may therefore be different.
Furthermore, the same station may change its measuring equipment over time. Contracts for station equipment are renewed periodically, but renewal does not guarantee that the same equipment will be maintained. All these features are shown in Figure 2.1 where concentration values of the herbicide prosulfocarb are used to illustrate the censoring.
Several methods have been developed to handle this type of data. We can cite the imputation of values to replace the LOQ values in the set of concentration values, the use of the maximum likelihood estimator or the Kaplan-Meier estimator (Gillaizeau et al., 2020; Croghan & Egeghy, 2003). In Chapter 4, we will present the method used for this type of data in the present thesis.

Irregular sampling

The second feature is a direct consequence of Section 2.1. We have already mentioned that each country has organised its own monitoring system in different ways. Some features are country-specific and have an impact on the raw data of the collected samples. In France, samples are

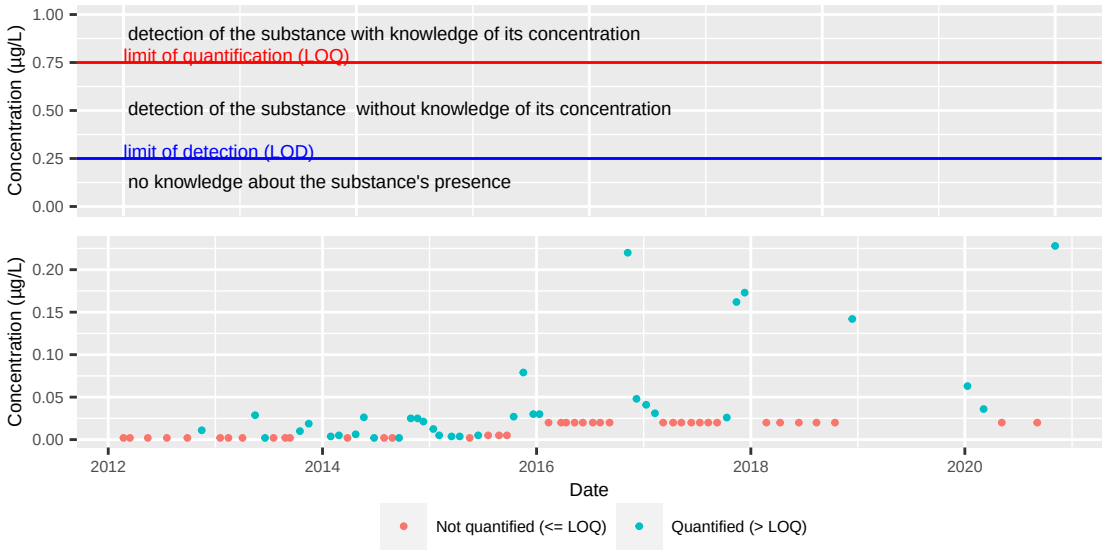

Figure 2.1: Censoring illustration. The top panel illustrates the limits of measurements effects. The bottom panel shows the consequences of censoring on the samples of a station located in the Centre-Val de Loire region. This station replaced its equipment in 2016, so the LOQ has changed. Note that the LOD and LOQ on the top panel are a theoretical example, which does not correspond to the real LOD and LOQ values of the station presented in bottom panel.

not necessarily collected at regular time steps. For instance, in pesticide data presented in Figure 2.1, no concentration data were collected in 2019.
Irregular sampling in time also happens in water monitoring data. This is a specificity of the French network, as this is not the case in all countries. For example, Zhang et al. (2008) states that sampling in surface waters is regularly carried out in China. Furthermore, Jørgensen & Stockmarr (2008) states that this is also the case for groundwater monitoring systems in countries such as China, the Netherlands and South Korea. The South Korean monitoring system is even automatic. Every six hours a sample is taken and after analysis it is automatically stored on a central server.
We would also like to point out that the application of other strategies in addition to regular sampling, such as grab sampling (Novic et al., 2017), could explain the irregular sampling rhythm in French surface waters. Grab sampling is about getting a picture of surface water quality over a limited area in a short time with highest possible accuracy. If grab samples and periodic monitoring of a station are recorded in the same database, irregular sampling may also occur.

Spatio-temporal heterogeneity in sampling

Another important feature mentioned in Baran et al. (2022) is the spatial and temporal heterogeneity of the records. This heterogeneity becomes clear when comparing measurements made in two different areas. Figure 2.2 illustrates that the activity periods of stations in two different geographical zones differ. Figure 2.2 also shows that the stations make very few measurements. The two groups have the same number of stations and yet group 2 has taken fewer

measurements than group 1 over the same period of time.

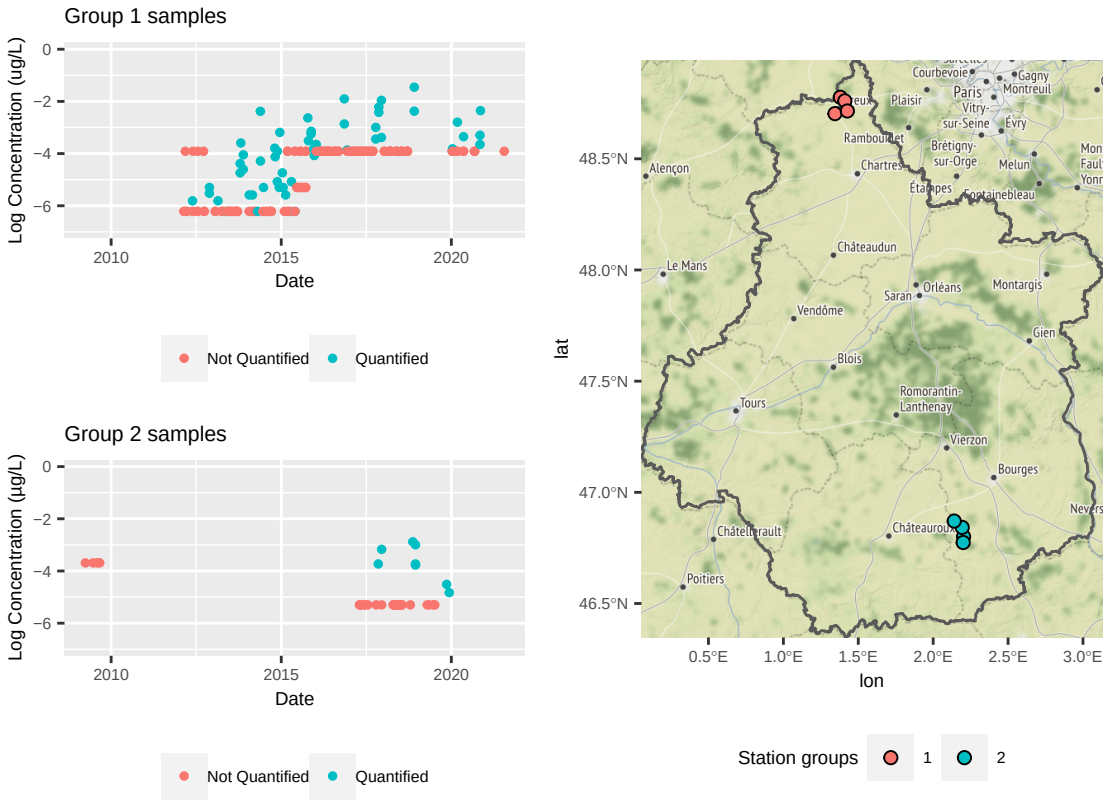

Figure 2.2: Spatial and temporal heterogeneity in sampling. The Figures on the left represent all the samples of two groups of stations. The map on the right shows the position of those groups. Concentration of prosulfocarb in surface waters are used for this illustration.

This problem of heterogeneity persists even with two spatially adjacent stations. In particular, measurements are not necessarily synchronous in time. In Figure 2.3 we see that it is difficult to compare the measurements of two existing stations. This would mean making strong assumptions about the temporal evolution of the concentrations since the stations performed sampling at non-overlapping time periods. One may notice that aggregating the data from the two stations results in a time series that is more evenly sampled over time. More details on this issue are provided in Section 2.4.

Spatio-temporal heterogeneity in analytical results

Figure 2.4 illustrates that, in addition to the temporal heterogeneity caused by the different sampling rhythms of the stations, the spatio-temporal data are not homogeneously distributed over the territory. The concentration values seem to differ completely across space. Figure 2.4 also illustrates that the distribution of concentration values can change drastically over time. Looking at the samples of station 03189000, there is a breakpoint in the concentration values of the station just before the year 2015. The same is true for the year 2016 in Figure 2.1.

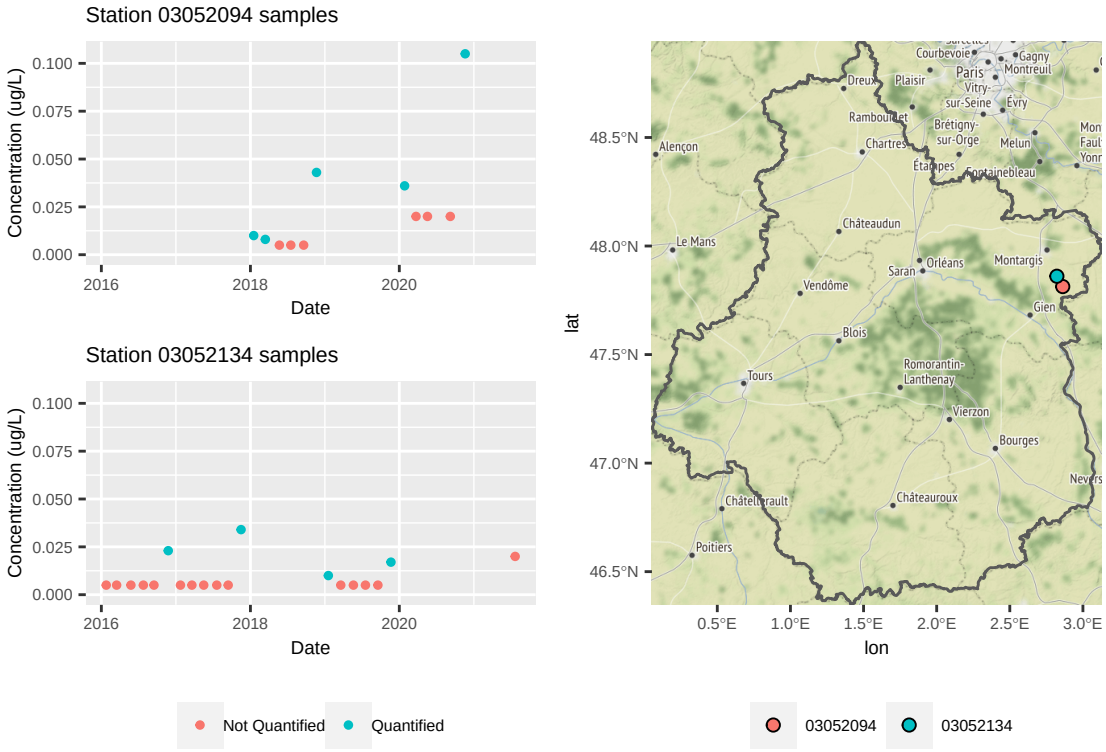

Figure 2.3: Spatial and temporal heterogeneity in sampling. The Figures on the left represent all the samples of two neighbouring stations. The map on the right shows the position of those stations. Prosulfocarb concentration in surface waters is used for this illustration.

This indicates that the use of a particular substance is not evenly distributed over the entire

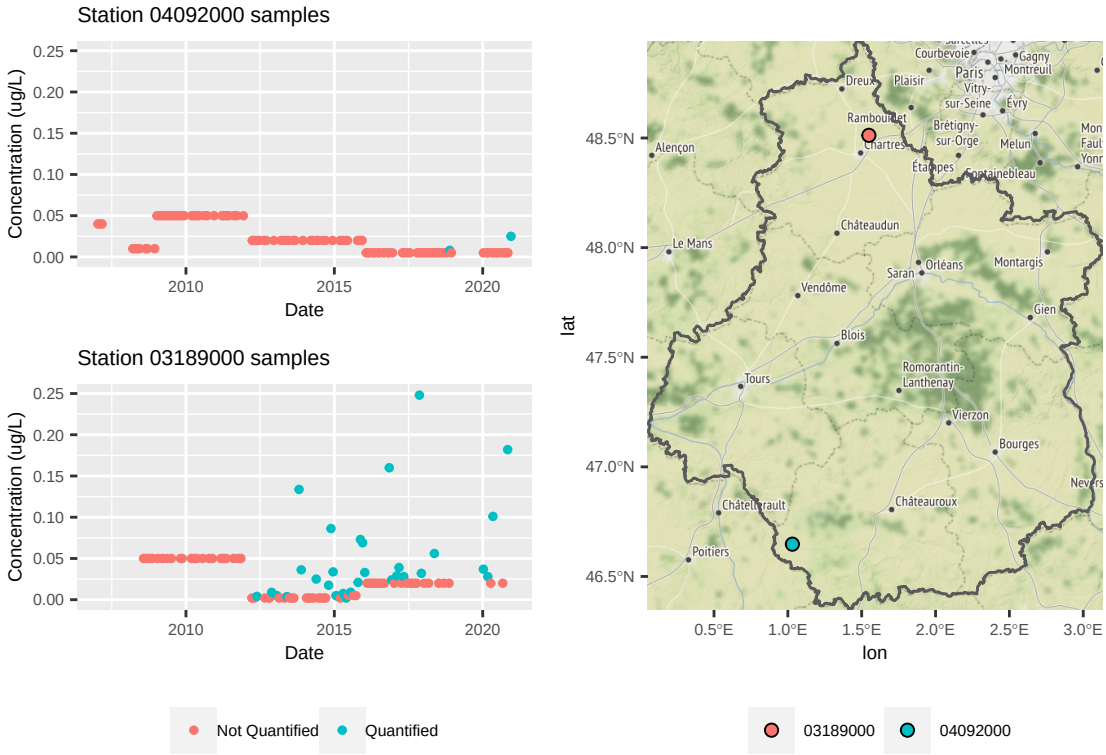

Figure 2.4: Spatial and temporal heterogeneity in distribution. The Figures on the left represent all the samples of two stations. The map on the right shows the position of those stations. Prosulfocarb concentration in surface waters is used for this illustration.

area. The spatio-temporal heterogeneity of concentration values can be easily explained by differences in agricultural practices depending on various factors related to professional habits and geological characteristics.

2.3.2. Indirect measurements

Surveys on farming practices

Additional indications of the presence of the substance are provided by surveys of agricultural practices. They are conducted by the Ministry of Agriculture in the form of a questionnaire and are used to describe and characterise how farmers work on their land. These surveys are very specific and focus only on certain types of crops. Three topics are addressed in the questionnaires. The first topic captures general information about the farm, such as a commitment to reducing pesticide use or agroecology. The second part of the questionnaire is designed to cover all technical operations on the farm plot. In other words, the department

of the ministry examines the structure of the plantation, its preceding crops or its irrigation. Finally, the use of pesticides on the entire farm is also examined. The questionnaire addresses criteria such as the type and settings of the sprayer for the substance or the handling and protection of the user. Figure 2.5 shows what kind of questions to expect in this survey [2]. In summary, surveys of agricultural practices 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 analyses by the statistical departments of the ministries of agriculture are publicly available (*Web resource: Agreste source*, n.d.), but not the raw data.

Opérations culturales

11 - Fertilisation et protection des cultures lors de la campagne 2019

Y-a-t-il eu au cours de cette campagne **au moins :**

- un apport de fumure organique ? ..... |__| 1 – oui |__| 0 – non
- un apport de fumure minérale ? ..... |__| 1 – oui |__| 0 – non
- un apport de fertilisant foliaire ? ..... |__| 1 – oui |__| 0 – non
- un traitement phytosanitaire (hors herbicides) ?..... |__| 1 – oui |__| 0 – non
- un apport de diffuseurs de phéromones ?..... |__| 1 – oui |__| 0 – non

Figure 2.5: Question extracted from the 2019 survey destined to viticulture. The question asked is about the fertilization and protection of the crops and the use of any phytosanitary product.

Substances sales databank

The use of a substance can also be seen indirectly in the sales data of plant protection products. The National Bank for the Sale of Pesticides by Authorized Distributors (NBSD) (Office français de la biodiversité and Système d'Information sur l'Eau, n.d.) lists and archives all such data. For reasons of anonymity, geographically fine resolution information is not available. The most precise resolution corresponds to postcodes. The same applies to the temporal resolution, which is no finer than the annual resolution. This dataset does not provide information on the location and date of use of the substance. In particular, a substance can be sold and used in distinct places. Nevertheless, sales give a general indication of the intensity of use of a substance. A sudden increase in sales of a product, as shown in Table 2.1, may mean that use in that area is increasing.

Crops cartography

Specific pest species can be observed for each crop type. Mapping the crop types in an area can therefore give a first idea of the areas and application periods of the substance to be monitored. Some of this information is available in the graphical land register (GLR) (*Web*

[2]Document in French, full document in *Web resource: Agreste* (n.d.)

	Year	Department	Substance	Quantity sold (in kg)	Annual rank
1	2008	INDRE	2,4-db	27.00	155
2	2009	INDRE	2,4-db	24.00	162
3	2010	INDRE	2,4-db	24.00	166
4	2011	INDRE	2,4-db	68.00	148
5	2012	INDRE	2,4-db	7.00	195
6	2013	INDRE	2,4-db	72.00	157
7	2014	INDRE	2,4-db	120.00	125
8	2015	INDRE	2,4-db	84.00	146
9	2016	INDRE	2,4-db	195.00	108
10	2017	INDRE	2,4-db	348.00	105

Table 2.1: Annual sales of the weed killer 2,4-db in the Indre department. The last column indicates the national annual rank of the substance sales.

resource: IGN data., n.d.). This database corresponds to the application forms used by farmers to obtain grants from the Common Agricultural Policy of the European Union (CAP). To be eligible for these subsidies, the crops grown on the plots must be declared. This dataset is partial information, as this declaration is compulsory only when applying for CAP subsidies. Therefore, the owners of the cultivated areas who have not applied for aid are not included in the database. Furthermore, 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. As an example, cultivation maps for barley and wheat are shown in Figure 5.13.

Adverse effects databases

The final example of data that can provide clues to the use of a substance are the databases used to monitor possible adverse events. They consist of medical registers that provide information on human and animal health. For human health, several sources of information can be cited:

- 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 they have any health problems. This organisation collects data through spontaneous reports from agricultural actors or during planned interviews with nurses or doctors.

- The medical-administrative databases of the MSA. They collect information on reimbursement for farmers' health care.

- The poison control centres are involved in monitoring adverse effects. They provide information on toxicovigilance for the whole population. Much information about acute health problems comes through these information channels.

- The National network of vigilance and prevention of professional pathologies (RNV3P), whose task is to identify emerging or re-emerging occupational health risks, is a good source for chronic health problems.

- The AGRICAN cohort (AGRIculture and CANcer) of the François Baclesse Centre is used to measure the health status of the agricultural population compared to the general population (especially in relation to cancer).

Regarding animal health, INRAE provides a database on veterinary toxicovigilance (GIS Toxinelle), and the Biodiversity French Office (OFB) is in charge of wildlife toxicovigilance. The Ministry of Agriculture provides additional information, e.g. on acute mortality in bees, and its biovigilance programme 500 ENI is also part of the available databases. This is a programme to monitor the impact of agricultural practices on biodiversity.

2.4. Substance diffusion-related data

We discussed the exposure factor in monitoring a health risk in Section 2.2. Exposure includes the ways in which the population may come into contact with the substance. The environment has a major influence on how exposure can occur. Therefore, it is important to include environmental information in the monitoring system. As an exhaustive list of all possible data sets would be too lengthy, this section provides examples of interesting additional data sets for surface water and air quality monitoring.

2.4.1. Surface waters quality

When monitoring the water quality of surface waters, stations are positioned at watercourses or lakes. GPS position of watercourses can provide information on how a substance might spread once it has entered the surface water system. This information is provided by the National Institute of Geographic and Forest Information (IGN) in the BDTOPO database (*Web resource: BDTOPO database*, n.d.). Figure 2.6 shows the river system of the French region Centre-Val de Loire with the positions of all stations. With this information, the concentrations of the stations can be directly compared according to their distance in the river system.

However, we mentioned in Section 2.3 that a single station does not provide many measurements. Therefore, in order to derive information from these data, one can work with coarser spatial grain. So there is a trade-off between the number of data available to make a statistical statement about a spatial area and the accuracy of the spatial resolution. Another interesting level of resolution, briefly mentioned in Section 2.2, is defined by the hydro-ecoregions (HER). These are geographical units in which hydrographic ecosystems share common characteristics. The criteria by which they are delineated combine features of geology, terrain and climate (Wasson et al., 2002). The INRAE services provide such information. Pooling all the samples from the HERs may provide a satisfactory level of aggregation, but this would be at the expense of spatial resolution. Figure 2.6 shows how the stations of the Centre-Val de Loire are distributed among HERs.

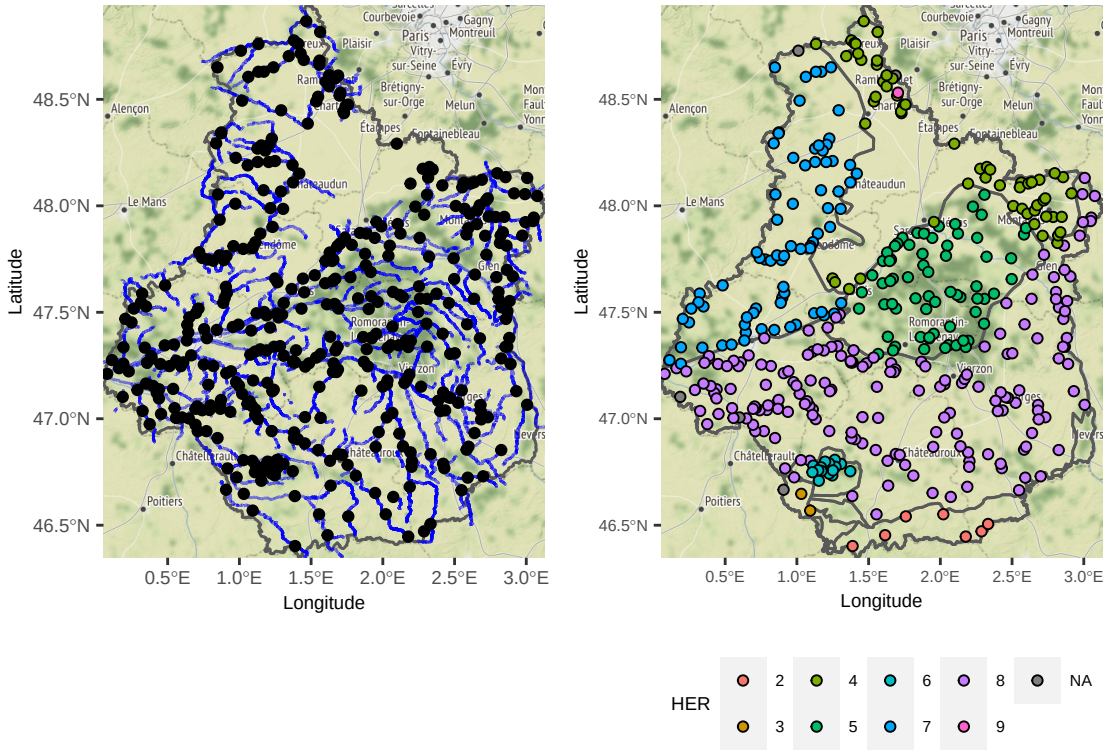

Figure 2.6: Stations monitoring surface waters in the Centre-Val de Loire French region. Two different geographical resolutions are represented. The resolutions are at station level in the left Figure and at HER level in the right Figure. The underlying hydrographic network linking all stations is plotted on the left, the stations are coloured according to their hydro-ecoregion on the right.

2.4.2. Air quality

Meteorological data is an important dataset for monitoring air quality, it provides for example information about the wind (wind direction, wind strength, etc.). Historical weather records are now available as open data on the Météo France website *Web resource: Météo-France data (SYNOP)*. (n.d.). Figure 2.7 illustrates the cross-referencing of data from air quality monitoring stations with meteorological data. Note that in this example we have chosen stations within the study region, but information from remote areas can also provide interesting information about air pollution in the selected area. This example also shows that concentration monitoring tasks depend on the application context. The coherence of each dataset included in the analysis needs to be discussed.

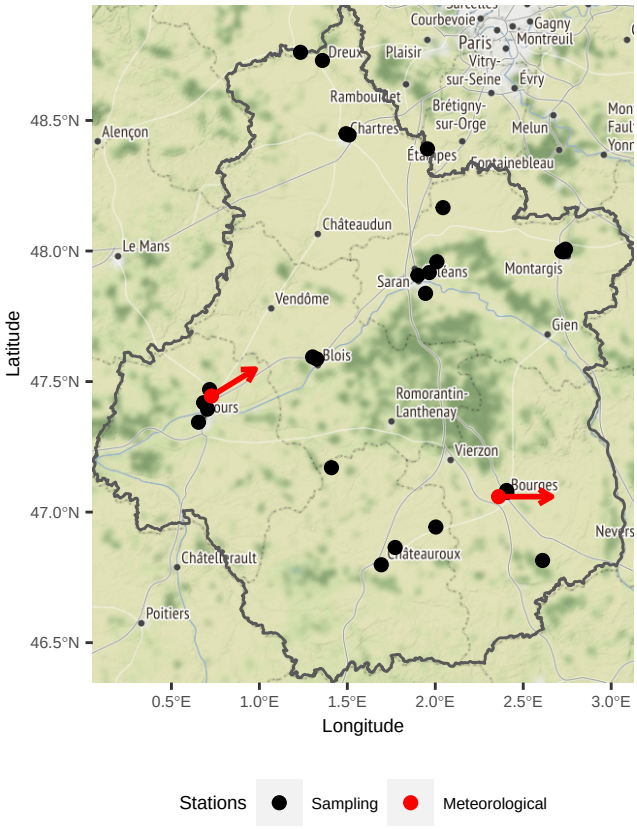

Figure 2.7: All active stations measuring air quality on March the 1st of 2021 coupled with meteorological stations active that day. The main wind direction and speed measured that day is mapped with the red arrows.

2.5. Surveillance of pesticides data

This section presents a first exploratory approach to monitoring pesticide concentrations using data from Sections 2.3 and 2.4. The spatio-temporal nature of concentration data requires appropriate techniques for their representation, which were reviewed in Andrienko et al. (2003); Cressie & Wikle (2015); Maimon & Rokach (2010). This includes the development of visualisation tools. There are several methods for visualising spatio-temporal data, but as mentioned in Ansari et al. (2019), spatial and temporal resolution is a key factor in the analysis and cannot be chosen automatically. Performing proper analysis of concentration data requires the involvement of experts. In this section, some visualisation techniques are presented and their limitations are discussed.

The spatial map plots or iterative maps (Andrienko et al., 2003) are an effective method to extract information from these data. They consist of maps of the same phenomenon at different times, as in Figure 2.8. There is a clear seasonal pattern in this figure. We can conclude that prosulfocarb is applied in autumn, and this information is affirmed by the crops targeted by this substance, namely winter wheat crops. The two years of observation were segmented by

the choice of temporal resolution of the seasons. Although this is a coherent choice, it has some limitations. For example, it cannot take into account years when the treatment started earlier or later due to climatic conditions. In addition, it cannot help to determine precisely the nature of the temporal change in the signal. Only the summarised indicator of quantification rate is used. Nothing is known about the maximum or average concentrations.
Another conventional representation is to display information with a map animation (Andrienko et al., 2003). In this method, the information displayed on the computer screen is updated according to the selected spatial area. Figure 2.9 shows a practical example of this technique. The concentrations of prosulfocarbe in the French region Centre-Val de Loire are displayed according to the selected HER region. There are two limitations to this display. The first one is the choice of spatial resolution. The HER were chosen to cluster the geography of the region. However, it can be seen that these can be very large regions. Stations located at opposite corners of a single HER may not have similar concentration values. Spatial heterogeneity of agricultural practices may occur at finer resolution. HER may not accurately capture regions where concentrations are homogeneously distributed. We will show how we deal with the spatial resolution issue in Chapter 5. This presentation also raises the question of the choice of temporal resolution. All samples available in the studied period are shown in Figure 2.9. There is a clear break in the three series around 2015. There seems to be a change in the concentration regimes.

In this last example, a finer temporal resolution combined with a comparison of the selected spatial regions could help experts to better understand the evolution of concentrations in the region. The aim of this thesis is to develop a statistical method that helps in the choice of both temporal and spatial resolution and also performs a comparison of the geographical regions at a given time.

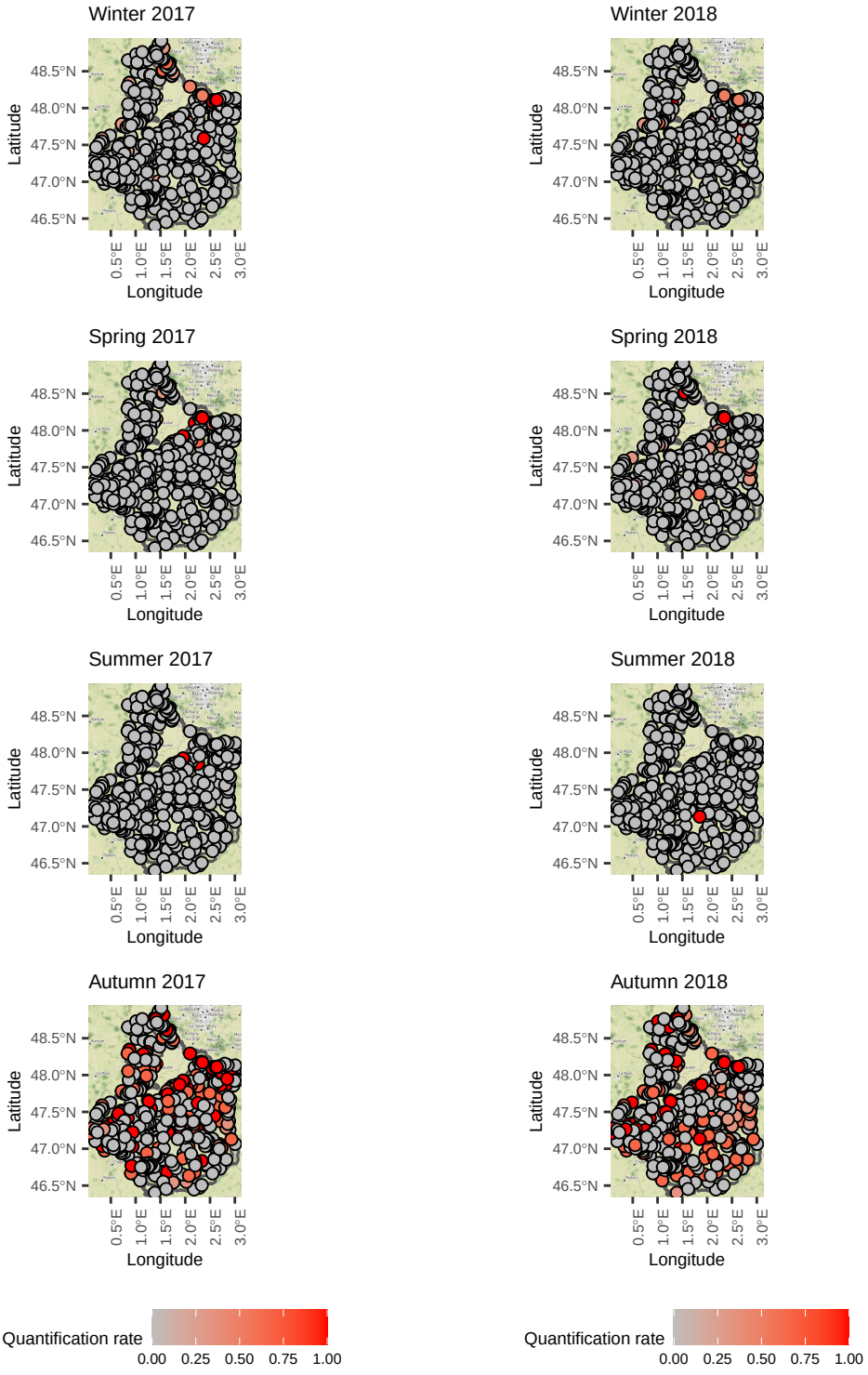

Figure 2.8: Spatial maps in time. Prosulfocarbe's quantification rate of each station was computed for each season of 2017 and 2018.

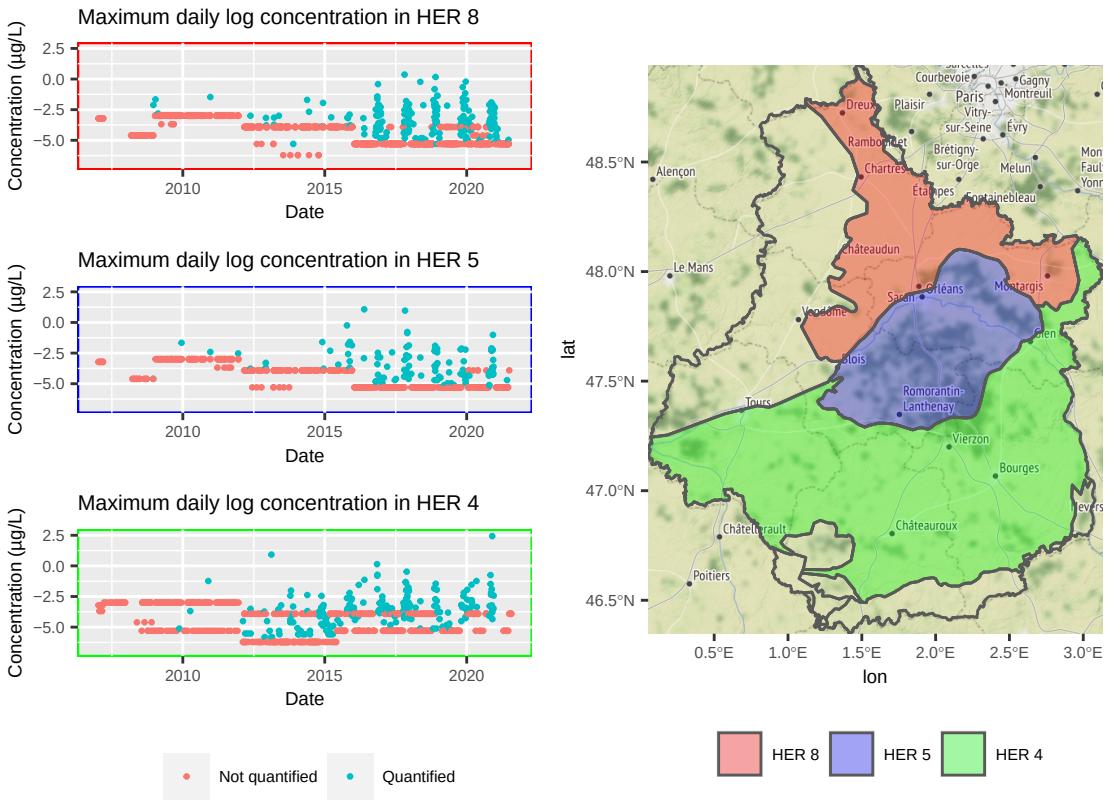

Figure 2.9: Time series plot of HER 8, 5 and 4 daily maximum concentrations. The log scale was used for a easier visualization.

2.6. Chapter summary

This chapter describes the tasks of the ANSES and the means by which it can fulfil them. The focus is on the task of phytopharmacovigilance. This is a system for monitoring pesticide concentrations and possible adverse effects on living organisms and the environment. The data needed to obtain information on such phenomena are not collected directly by the Agency, but are provided by various administrative departments. Chemical concentration data have specific characteristics such as censoring due to the accuracy of the measuring instruments, spatial and temporal heterogeneity of sampling due to the specificities of the French monitoring network, or spatial and temporal heterogeneity of concentration values explained by the different agricultural practices across the territory. Other datasets can be consulted to get a better understanding of concentration levels and the distribution of the substances. In practice, one way to monitor substances is to combine all these sources of information. Some monitoring tools use visualisation techniques based on these data sets. Although each visualisation technique makes it possible to obtain results in the monitoring mission, it also has disadvantages. We therefore propose to combine visualisation and modelling to improve the monitoring of substances.
First, the global spatio-temporal heterogeneity of the concentration data must be remedied by searching for homogeneous time periods and geographical areas. We will first focus on the search for homogeneity in the temporal aspect. In Chapter 3, we review the literature to find suitable methods for this task. We then adapt these methods to the specifics of the concentration data in Chapter 4. Once the method is selected, we describe a procedure to deal with spatial heterogeneity and identify abnormal signals of concentration in Chapter 5. In Chapter 6, we present a monitoring tool that combines statistical inference and visualisation method.

3. A selection of methods for change point detection

Contents

Let us recall that our first objective is the detection of homogeneous temporal periods in the use of a chemical substance for which monitoring data is available. This task can be viewed as a change point detection problem, where the change points delineate homogeneous time intervals. In this chapter, we propose to go over some available methods to solve this problem.

We have chosen to limit this chapter to off-line methods, i.e. methods that search for changes in known features of fully acquired signals. This choice is motivated by the speed of data collection and storage in the context of pesticide concentration monitoring, as these data are not collected in real time. Nevertheless, for the sake of completeness, we can mention that there exist quite a number of on-line detection methods in the literature (S. Liu et al., 2017; Y. Li et al., 2021; Höhle, 2010; Ranganathan, 2010; S. Li et al., 2015).This present thesis draws on the following research (Truong et al., 2020; Basseville & Nikiforov, 1993; Bardet, Jean-Marc et al., 2020).

This chapter is structured as follows: Section 3.1 presents a classical modelling framework for change point detection and explores the different ways to evaluate temporal sub-signals; Section 3.2 presents existing methods for estimating the positions of a known number of changes; Section 3.3 reviews a selection of methods that have been proposed to estimate the number of change points when the latter is unknown; Section 3.4 reviews efficient available algorithms to obtain different segmentation results; Section 3.5 looks at applying change point detection methods to data that have the same characteristics as in Section 2.3.

3.1. Model and cost functions

We describe the most general framework for a change-point model. We consider a signal consisting of observations $\boldsymbol{y} = (y_1, ..., y_n)$, which are the realisations of random variables $Y_1, ..., Y_n$. The variables Y_i are recorded sequentially, and the recording times are not necessarily equidistant. Integrating the irregular sampling times or the temporal gaps between the observations in the modelling is not in the scope of this survey. Thus, the indices in Y_i are only indicators of the order of occurrence in the sample and not of the observation times.

Some characteristics (trend, mean, variance, etc...) of the signal $\boldsymbol{y}$ are supposed to change at the K^* indices $\tau_1^* < ... < \tau_k^* < ... < \tau_{K^*}^*$. Moreover, we set $\tau_0^* = 0$ and $\tau_{K^*+1}^* = n$. The purpose of breakpoint detection is to estimate the positions τ_k^* and the number of breaks K^* when they are unknown. We denote $y_{u:v}$ as a segment of the signal from the u-th coordinate to the v-th. The goal is to identify such segments in the data for which the characteristics mentioned above are stable.

According to the nomenclature proposed by Truong et al. (2020), change point detection methods are based on a cost function W for individual segments. Intuitively, when the properties (on which changes are investigated) of the segment $y_{u:v}$ are homogeneous, the cost $W(y_{u:v})$ takes low values. Supposing that the segments are independent, the total cost $\mathcal{C}(\boldsymbol{y}, \mathcal{T})$ associated with a segmentation $\mathcal{T} = (\tau_k)_{k=1}^K$ with K change points is given as the sum of the costs of all segments:

$$\mathcal{C}(\boldsymbol{y}, \mathcal{T}) = \sum_{k=0}^{|\mathcal{T}|} W(y_{\tau_k+1:\tau_{k+1}}), \tag{3.1}$$

With these notations and assuming K^* is known, finding the change points is equivalent to solving the optimization problem:

$$\widehat{\mathcal{T}} = \arg\min_{|\mathcal{T}|=K^*} \mathcal{C}(\boldsymbol{y}, \mathcal{T}) = \arg\min_{|\mathcal{T}|=K^*} \sum_{k=0}^{K^*} W(y_{\tau_k+1:\tau_{k+1}}) \tag{3.2}$$

The choice of the cost function W determines the type of changes (in trend, mean, etc.) targeted by the detection. Figure 3.1 illustrates two different types of changes that may be of interest for change point detection. In the next parts of this section, we distinguish the cost functions according to the statistical inferences which they are based on, namely parametric or non-parametric inference. We give a non-exhaustive list of cost functions for each inference framework.

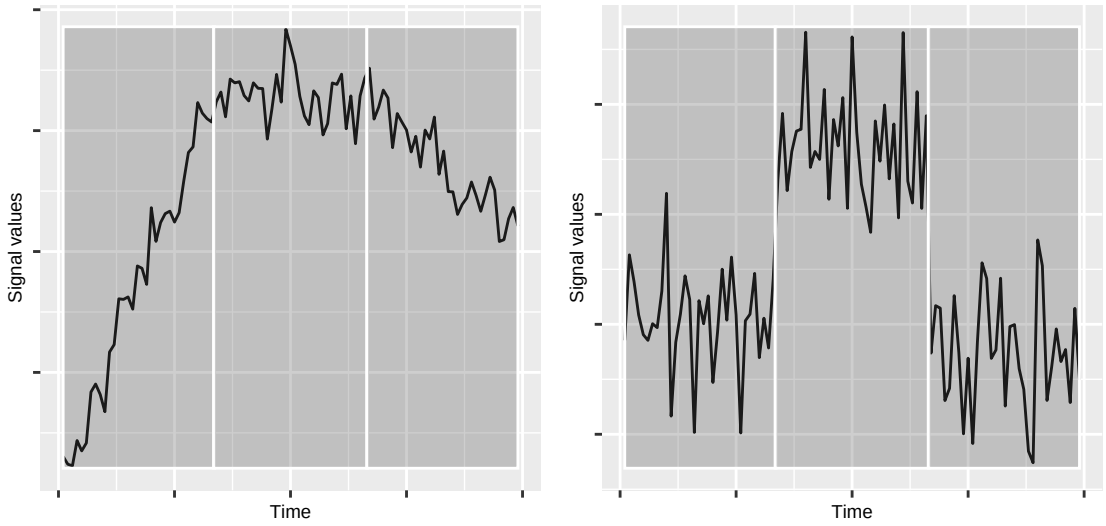

Figure 3.1: Examples of types of change point detection.
The figure on the left illustrates changes in trend: the data is simulated using a linear trend with a noise term following a normal distribution with known variance.
The figure on the right illustrates changes in mean: the data follow a normal distribution with known variance and changes in mean.

3.1.1. Parametric inference

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 (Bai, 1994; Fearnhead et al., 2018) does not require the same modelling as detecting changes in the mean (Frick et al., 2014; J. Chen & Gupta, 2012). We highlight these differences below.
A first classical cost function is based on the maximum likelihood estimation. In this setting, it is assumed that the observations located in the k-th segment follow a distribution Q depending on a vector of parameters θ_k^* with $\theta_k \in \Theta$, a compact subset of $\mathbb{R}^p$. In other words, we suppose

that all observations are sampled from the same distribution Q but the values of θ_k^* change abruptly at each change-point τ_k^*. More formally, we have that:

$$y_t \sim \sum_{k=0}^{K^*} f(.;\theta_k^*)\mathbb{1}_{\tau_k^*+1\leq t\leq \tau_{k+1}^*}, \tag{3.3}$$

with f being the density function of distribution Q and the observed signal $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ is composed of independent random variables. The cost function used to evaluate the segments in this context is the negative log-likelihood calculated using $\hat{\theta}_{ML}$, the maximum likelihood estimator of the segment parameter. Hence, for a segment $y_{u:v}$ with $u < v$, we can write:

$$W(y_{u:v}) = -\sup_{\boldsymbol{\theta}\in\Theta} \sum_{i=u}^{v} \ln f(y_i;\boldsymbol{\theta}) \tag{3.4}$$

This method would prove useful in the example presented in the right side of Figure 3.1. Applying the maximum likelihood estimator on the mean of Gaussian distribution would provide satisfying results. Other distributions than the Gaussian were investigated (Maidstone, 2016; Frick et al., 2014) since it is not always well suited for data (especially concentrations data).
Cost functions adapted for changes in trend rely on piecewise linear regression. We place ourselves in the simplest case where $\boldsymbol{y}$ is univariate response to observed covariates $\{x_t\}_{t=1}^n$ such that $x_t \in \mathbb{R}^p$. Observations located in the k-th segment are modelled as:

$$y_t \sim \sum_{k=0}^{K^*} (x_t'\theta_k^* + \epsilon_t)\mathbb{1}_{\tau_k^*+1\leq t\leq \tau_{k+1}^*}, \tag{3.5}$$

where $\theta_k^* \in \mathbb{R}^p$ are the regression parameters and ϵ_t is the noise on the signal, assumed to be Gaussian with zero mean and known variance in this precise setting. The adapted cost function in this configuration uses the least squares estimation and is expressed as:

$$W(y_{u:v}) = \min_{\theta\in\mathbb{R}^p} \sum_{t=u}^{v} (y_t - x_t'\theta)^2 \tag{3.6}$$

This modelling is perfectly suited for the detection of the changes in the left example of Figure 3.1. Note that it can be straightforwardly extended to signals with non gaussian noise.

3.1.2. Non-parametric inference

Cost functions for segments have also been formulated for the non-parametric framework. Several strategies have been developed in the literature over time. These include the non-parametric maximum likelihood method (Zou et al., 2014; Einmahl & McKeague, 2003), kernel methods (Harchaoui et al., 2008; S. Li et al., 2015), and rank-based methods (Pettitt, 1980; Wang et al., 2019). We will focus on the latter because it was adapted for censored observations in Lung-Yut-Fong et al. (2015).
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 $R_i =$

$\sum_{j=1}^{n} \mathbb{1}(Y_j < Y_i)$. Moreover, we note $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i < t)$ the empirical cumulative distribution function (c.d.f.). The cost function is derived from the Wilcoxon/Mann-Whitney rank criterion. Indeed, finding a single breakpoint in the signal amounts to testing the following assumptions:

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

For this, Lung-Yut-Fong et al. (2015) introduce a test statistic for $\mathcal{H}_0$ and $\mathcal{H}_1$ defined as:

$$S_n(t) = \hat{\Sigma}_n^{-1} U_n^2(t), \tag{3.7}$$

where $U_n(t)$ is the centred rank statistic of the t-th observation, expressed as:

$$U_n(t) = \frac{2}{\sqrt{nt(n-t)}} \sum_{i=1}^{t} \left(\frac{n+1}{2} - R_i \right), \tag{3.8}$$

and $\hat{\Sigma}_n = \frac{4}{n}\sum_{i=1}^{n}(\hat{F}_n(Y_i) - 1/2)^2$. Theorem 1 of Lung-Yut-Fong et al. (2015) 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 change point detection by Lung-Yut-Fong et al. (2015). The cost function W for a segment $y_{u:v}$ is defined as:

$$W(y_{u:v}) = -(v-u)\hat{\Sigma}_n^{-1}\overline{R}_{u:v}^2, \tag{3.9}$$

where $\overline{R}_{u:v} = \frac{1}{v-u}\sum_{i=u}^{v} R_i$ is the average rank of $y_{u:v}$. This method identifies segments where the ranks of the observations are homogeneous. It would be very efficient in the case of mean change detection as presented in the right panel of Figure 3.1.
It is also possible to derive change detection in trend using non parametric inference. The experiments of Haynes et al. (2016) show that a non-parametric likelihood method finds results similar to those obtained with a piece-wise regression change-point model. The Mann-Kendall test statistic (Pohlert, 2020; "Chapter 23 Nonparametric Tests for Trend Detection", 1994) seems to be also a good candidate cost function to derive a detection method for changes in trend.

3.2. Optimal partition search method

In this section, we turn to change point location estimation. Various search methods for finding change points have been described in the literature. They can be distinguished according to whether they provide an optimal solution to the problem of the search of change points or an answer in the form of an approximation. Approximation methods are not discussed, but there are plenty of them, such as sliding window methods (W. Li et al., 2010; C. Liu et al., 2022), bottom-up segmentation (S. Chen et al., 1998), and binary segmentation (Yang & Kuo, 2001; Fryzlewicz, 2014).

We choose to focus on the optimal partition method. This choice is motivated by the size of the datasets we apply change-point detection on. The number of samples is still in a reasonable range to obtain satisfying computational times.
The optimal segmentation algorithm computes an exact solution to the problem defined in Equation (3.2) (Auger & Lawrence, 1989). This search method is based on dynamic programming. It requires to compute the cost of all possible segments $y_{u:v}$ with $u < v$. With a fixed K_{max} number of change-points, one can recursively solve the optimization problem. The recursion comes from the following relationship:

$$\min_{|\mathcal{T}|=K_{max}} \mathcal{C}(\boldsymbol{y}, \mathcal{T}) = \min_{t \leq n-K_{max}} \{W(y_{1:t}) + \min_{|\mathcal{T}|=K_{max}-1} \mathcal{C}(y_{t+1:n}, \mathcal{T})\} \tag{3.10}$$

In other words, given all possible segmentations of all sub-signals $y_{t:n}$ in $K_{max}-1$ segments, one can compute the optimal segmentation of the whole signal $\boldsymbol{y}$ in K_{max} segments. This results in a computational cost of order $\mathcal{O}(K_{max}n^2)$ (Haynes et al., 2017). Algorithm 1 provides the implementation of (3.10).

Algorithm 1 Optimal partition algorithm:

input : signal $y_{1:n}$, cost function $W()$, number of change points $K_{max} \geq 1$
Create C_1 a $n \times n$ empty matrix
for all (u, v) such that $1 \leq u < v \leq n$ **do**
    $C_1(u, v) \leftarrow W(y_{u:v})$
end for
if $K_{max} + 1 > 2$ **then**
    for $k = 2, ..., K_{max}$ **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 $K_{max} + 1$
$L_{K_{max}+1} \leftarrow n$
$k \leftarrow K_{max} + 1$
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_{max} estimated change points (with n as a last coordinate).

Notice that slight modifications can be made to Algorithm 1 to obtain the output of all segmentation results for all values of $K \leq K_{max}$.
The downsides of the optimal partition method reside in its computational cost which is expensive and in the fact that problem (3.2) suppose that the number of changes K^* is known.

There are ways to compute an estimate of K^* from the results of optimal partitioning. We will address this issue in detail in the next section.
Different applications of optimal partitioning methods can be found in Rigaill (2015); Lavielle (1999); Perron et al. (2006). We will see in Section 3.4 another optimal approach in a different setting involving a penalized criterion.

3.3. Estimation of an unknown number of changes

This section discusses the case where the number of change points K^* is unknown. This can be justified in our application context the following way. Since pesticides are supposed to be spread at regular times during the years, we could expect a seasonal behaviour of their concentrations and thus have a clue on the number of changes before-hand. However, the spatio-temporal heterogeneity discussed in Section 2.3 prevents us from being certain of this assumption. Additionally, our ultimate goal is to help detecting anomalies, which could precisely be an abnormal spread of a substance in the environment outside the expected time periods. This is a second reason to consider that the number of change points is unknown.
Choosing the optimal number of change points can be seen as a model selection problem. The difficulty in the choice of K^* is that the cost of a segmentation decreases when the number of breakpoints increases. Thus, the choice of a high value of K^* can lead to an overfitting model. We need to find methods providing a segmentation of the signal into an acceptable number of segments e.g. one that lower the cost enough but does not overfit to the data too much.
We give two examples of how to proceed.

- **Using an elbow heuristic:** this heuristic provides an estimate of K^* without involving a penalization procedure and is notably used in Lung-Yut-Fong et al. (2015). It is based on the plot of the costs with respect to their number of change points. It consists in fitting the best two part linear model on the costs $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_0), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}})\}$. In other words, $\widehat{K}$ is the number of change-points K that minimizes the residual sum of squares of the two linear models fitted on $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_0), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_K)\}$ and $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_K), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}})\}$. Algorithm 2 illustrates the elbow method procedure.

- **Penalizing the cost:** as mentioned in Truong et al. (2020), the optimization problem (3.2) can be modified when the number of breaks is unknown by adding a penalty term. Intuitively, the penalty term acts as an additional cost one must pay each time a break is decided in the signal $\boldsymbol{y}$. This gives the new optimization problem:

$$\min_{\mathcal{T}} \{\mathcal{C}(\boldsymbol{y}, \mathcal{T}) + pen(\mathcal{T})\} \tag{3.11}$$

  Once the optimal partitioning method has been applied with a maximum number of change points K_{max}, we obtain the resulting segmentations $\{\widehat{\mathcal{T}}_0, \ldots, \widehat{\mathcal{T}}_{K_{max}}\}$ and their associated costs $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_0), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}})\}$. We can apply the penalization procedure on these costs and estimate K by selecting the minimal penalized cost:

$$\widehat{K} = \arg \min_{K \in \{0, \ldots, K_{max}\}} \{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_1) + pen(\widehat{\mathcal{T}}_1), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}}) + pen(\widehat{\mathcal{T}}_{K_{max}})\} \tag{3.12}$$

Algorithm 2 Elbow method algorithm

input : the segmentations cost resulting from optimal partitioning $\mathcal{C}(\boldsymbol{y}, \mathcal{T}_K)$ for $K \in \{0, ..., K_{max}\}$

initialisations : Initialize $C \leftarrow (\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_0), ..., \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}}))$,
Initialize $slope \leftarrow (0, ..., 0)$ a $K_{max} - 2$ length vector.
for $k = 1, ..., K_{max} - 1$ **do**
  Compute $ml1$ the linear regression model that regresses $C(0 : k)$ on $(0 : k)$
  Compute $ml2$ the linear regression model that regresses $C(k : K_{max})$ on $(k : K_{max})$
  Set $slope(k-1)$ as the sum of residuals of $ml1$ plus the sum of residuals $ml2$.
end for
$CP \leftarrow \arg\min_{k \in \{1,...,K_{max}-1\}}(slope)$
output : the optimal number of changes CP.

$\widehat{K}$ and the segmentation $\mathcal{T}_{\widehat{K}}$ are the optimal solution for the change-point search when K^* is unknown. Several penalization strategies are presented in Truong et al. (2020).

Penalizing the costs is not only useful in the choice of an optimal number of change points K^* but also useful to design efficient algorithms. Under some assumptions described in the following section, these algorithms can in particular be more efficient than the optimal partition method presented in Algorithm 1.

3.4. Strategies using a penalized criterion

When using a penalized criterion (as the one of Equation (3.11)) in change point detection procedure, there are two separate topics that must be discussed. The first one is the search method when the penalty term is fixed, it is addressed in Section 3.4.1. The second topic is the efficient exploration of the penalty weights to obtain several segmentation results, it is addressed in Section 3.4.2.

3.4.1. PELT search method

Problem (3.11) can be solved with an efficient dynamic programming method under some specific penalization strategy. The Pruned Exact Linear Time (PELT) algorithm was introduced by Killick et al. (2012). It is efficient when the penalization strategy is linear in the number of change point K. More formally, the penalty term writes as:

$$pen(\mathcal{T}) = |\mathcal{T}|\beta$$

The penalty value (or weight) parameter β takes positive values. It corresponds to the cost assigned to a breakpoint. Intuitively, the number of change point detected is low when the penalty value is high. The number of change point is indeed a non increasing function of the penalty value.

The general ideas underlying the PELT method are the following:

- the signal is explored sequentially
- for each index $s \in 1, ..., n$:
  - a set of potential breakpoints $\{\tau_0, ..., \tau_m\}$ is obtained
  - a pruning rule involving the penalty value β is used to eliminate candidates from this set.

This is the principle of Algorithm 3. The pruning rule of Killick et al. (2012) can be stated as follows:
for all $t < s < n$, if

$$\min_{\mathcal{T}} \left[\mathcal{C}(y_{1:t}, \mathcal{T}) + |\mathcal{T}|\beta \right] + W(y_{t+1:s}) + \beta \geq \min_{\mathcal{T}} \left[\mathcal{C}(y_{1:s}, \mathcal{T}) + |\mathcal{T}|\beta \right], \tag{3.13}$$

holds, then t can never be the last change point prior to n.
We introduce the following additional notation to simplify the algorithm writing:

$$F(s) = \min_{\mathcal{T}} \left[\mathcal{C}(y_{1:s}, \mathcal{T}) + pen(\mathcal{T}) \right]$$

The notation $F(s)$ corresponds to the best partition possible of the sub-signal $y_{1:s}$.

Algorithm 3 PELT algorithm

input : the data $y_1, ..., y_n$, a cost function $W()$, the penalty term β and a minimal segment length n_{min}

initialisations : F a vector of size n, $R_1 = \{0\}$, $CP(0) = NULL$
$F(i) = -\beta$, for all $i \in \{1, ..., n_{min}\}$
for all $\tilde{t} = n_{min} + 1, ..., n$ **do** :
  Compute $F(\tilde{t}) = \min_{t \in R_{\tilde{t}} || t - \tilde{t}| \geq n_{min}} \{F(t) + W(y_{(t+1):\tilde{t}}) + \beta\}$
  Compute $\bar{t} = \arg\min_{t \in R_{\tilde{t}} || t - \tilde{t}| \geq n_{min}} \{F(t) + W(y_{(t+1):\tilde{t}}) + \beta\}$
  Set $CP(\tilde{t}) = [CP(\bar{t}), \bar{t}]$
  Set $R_{\tilde{t}+1} = \left\{ t \in R_{\tilde{t}} \cup \{\tilde{t}\} | F(t) + W(y_{(t+1):\tilde{t}}) + \beta \leq F(\tilde{t}) \right\}$
end for
output : the vector of change-points $CP(n)$.

In Algorithm 3, the parameter n_{min} is introduced to guarantee a minimum segment size. It has a direct influence on the segmentation and acts as a compromise between the segmentation resolution and the parameter estimation statistical validity. Small values of n_{min} can potentially lead to the detection of a large number of change-points. However, the statistical validity of these changes can be questioned. Large values of n_{min} ensure enough data for the parameter estimation, but there is a risk to miss changes that occurred at close index locations.

The complexity of PELT can reach $\mathcal{O}(n)$ when the change points are supposed to be distributed uniformly over the signal $\boldsymbol{y}$. This constitutes a major improvement compared to the optimal partitioning method. However, as mentioned in Haynes et al. (2016), the penalty value β has an influence on the performance of PELT.
For a single value of penalty, PELT returns a single segmentation of $\boldsymbol{y}$ (e.g. the result for a single value of K). However, we cannot predict the number of breakpoints resulting from a given fixed value of β. Thus, we do not know if the change point model resulting from the choice of β is over or under fitting the signal $\boldsymbol{y}$. We need an exploratory approach to tackle this issue. This is the objective of next section.

3.4.2. CROPS: exploring a range of penalty values

Diverse strategies to calibrate β exist as the BIC criterion which is widely used (Yao, 1988; Faure et al., 2016; Shi et al., 2022). One of the main drawbacks is that penalised criteria such as the BIC penalty are only valid under certain assumptions. These assumptions include exploring the true model, e.g. that the data is truly distributed under the assumed model. Since it is impossible to check this in practice, some data-driven calibrations for the penalty values have been developed (Birgé & Massart, 2006; Baudry et al., 2011; Bardet et al., 2012; Arlot & Massart, 2009). They were designed to select better suited penalty values for a given dataset.
However, if we reconnect this short review of methods to the application context of this manuscript, it would be interesting to present several segmentation results to the experts. In order to do so, we would like to find a compromise between the completeness of the solution provided by the optimal partition method (e.g. the segmentation results for all $K \leq K_{max}$) and the computational cost of PELT.
The algorithm CROPS (Change points for a Range Of PenaltieS algorithm Haynes et al. (2017)) is scanning a range of penalties $[\beta_{min}, \beta_{max}]$, and automatically focuses on the values at which new change points are introduced.
The process to uncover new penalty values $\beta \in [\beta_{min}, \beta_{max}]$ is based on theorem 3 of Haynes et al. (2017). Noting $U_K(\boldsymbol{y}) = \min_{|\mathcal{T}|} \mathcal{C}(\boldsymbol{y}, \mathcal{T})$ the unpenalized cost of the optimal segmentation in K change points of $\boldsymbol{y}$ and $K(\beta)$ the number of change points of the optimal segmentation result obtained using β in problem (3.11), this theorem writes as follows:

Theorem 3.4.1. *Let $\beta_0 < \beta_1$, 3 cases are possible to uncover new penalty values:*

1. *If $K(\beta_0) = K(\beta_1)$ then $K(\beta) = K(\beta_0)$ for all $\beta \in [\beta_0, \beta_1]$*
2. *If $K(\beta_0) = K(\beta_1) + 1$ then $K(\beta) = K(\beta_0)$ for all $\beta \in [\beta_0, \beta_{int}[$ and $K(\beta) = K(\beta_1)$ for all $\beta \in [\beta_{int}, \beta_1]$ with:*
$$\beta_{int} = \frac{U_{K(\beta_1)}(y_{1:n}) - U_{K(\beta_0)}(y_{1:n})}{K(\beta_0) - K(\beta_1)} \tag{3.14}$$
3. *If $K(\beta_0) > K(\beta_1) + 1$ and $K(\beta_{int}) = K(\beta_1)$ where β_{int} is defined in (3.14), then $K(\beta) = K(\beta_0)$ if $\beta \in [\beta_0, \beta_{int}[$ and $K(\beta) = K(\beta_1)$ if $\beta \in [\beta_{int}, \beta_1]$*

From this theorem, the authors propose the search Algorithm 4:

Algorithm 4 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 `PELT`$(y_{1:n}, \beta_{min})$ and `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 $K(\beta_0) > K(\beta_1) + 1$ **then**
    $\beta_{int} \leftarrow \frac{U_{K(\beta_1)}(y_{1:n}) - U_{K(\beta_0)}(y_{1:n})}{K(\beta_0) - K(\beta_1)}$
    $res \leftarrow$ `PELT`$(y_{1:n}, \beta_{int})$
    From res store $K(\beta_{int})$
    if $K(\beta_{int}) \neq K(\beta_1)$ **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 $\beta \in [\beta_{min}, \beta_{max}]$.

As stated in Haynes et al. (2017), the theoretical upper bound for the number of times PELT has to run to find all possible segmentations with $\beta \in [\beta_{min}, \beta_{max}]$ is given by $K(\beta_{min}) - K(\beta_{max}) + 1$. For each β uncovered by CROPS, we have the number of changes $K(\beta)$ and the cost $\mathcal{C}_\beta(\boldsymbol{y}_{1:n})$. We can then run Algorithm 2 to estimate an optimal number of changes and keep the results of all segmentations found by CROPS for exploratory purposes.
Note that we can finally select a segmentation by combining Algorithm 4 with the elbow method applied to the curve plotting $\mathcal{C}_\beta(\boldsymbol{y}_{1:n})$ as a function of $K(\beta)$ (for each β uncovered by CROPS).

3.5. Change-point detection in environmental data

Change point detection is a reference technique for time series segmentation and is used in a variety of applications (Basseville & Nikiforov, 1993; J. Chen & Gupta, 2012; S. Liu et al., 2017; Reeves et al., 2007; Lévy-Leduc & Roueff, 2009), in particular for environmental pollution monitoring (Costa et al., 2016).
The application domains of the studies we found in the literature are specific and they can significantly differ from the phytopharmacovigilance topic. Three aspects are recurrent in these studies:

- Change point detection is quite often performed on aggregated indicators with a yearly or monthly temporal resolution (Ko et al., 2017; Ryberg et al., 2020; Fomby & Lin, 2006). The temporal aggregation highlights the fact that most studies have to handle high irregular sampling in the raw data.

- Changes in trend are ubiquitous in the studies we found. Looking for changes in trends underlines that the observed phenomena have somewhat of a continuous behaviour in time or at least exhibit temporal correlation. Examples of applications are resistance appearance to a substance (de Solla et al., 2010), the evolution of the exposition of animal populations to pesticide (Menger et al., 2022) or impacts of public policies on air quality (Fomby & Lin, 2006).

- Optimal search methods are used in change point detection on environmental data (Bunce et al., 2018; Ryberg et al., 2020). Even though they are not the most common methods in the studies we found, they are still applicable to environmental data with consequent size.

Aggregating the data is a way to smooth the data and reduce the noise that could be present. Smoother data can justify the detection of changes in trend. This is an important point since we also aggregate concentration data in Chapter 5 to a daily resolution. We are now assured this is common practice in this type of applications.
However, aggregation with a daily resolution does not introduce much smoothing in the data. Moreover, the characteristics of concentration data are the censoring and the spatio-temporal heterogeneity (see Chapter 2). This is why we do not consider trend detection in our approach. Investigating further on the two aspects of censored data and spatio-temporal heterogeneity gave interesting results but this important characteristics of environmental data is rarely, not to say never, taken into account in environmental studies because the use of aggregated indicators is a way to circumvent this problem.
The second aspect to consider is spatio-temporal heterogeneity. Air pollution data have 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 (Sampson et al., 2011; Lindström et al., 2014), while others use latent variables and co-clustering approaches (Bouveyron et al., 2022).
Nevertheless, these approaches do not include the fact that the data under consideration are not normally distributed, and are 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 Ccanccapa et al. (2016); Figueiredo et al. (2021); Aznar et al. (2017).

3.6. Chapter summary

This Chapter reviews the change-point detection literature. The main statistical inferences are presented in Section 3.1. It is possible to detect changes in a signal by parametric or non parametric inferences and the choice of the cost function is determinant of the type of changes detected. Section 3.2 goes over methods to estimate an unknown number of change-points and their locations. Even though the optimal partitioning method is designed to detect a fixed number of changes, an estimate of the optimal number of changes can be derived from heuristics such as the elbow methods. Another strategy to compute an estimate is to penalize the costs resulting from the optimal partitioning methods. Under a specific penalization strategy, the PELT algorithm can be used. This results in a considerable speed up in computational time. However, doing so adds a a penalty parameter which is important to calibrate. Instead of setting the penalty to an optimal value, Section 3.3 presents a more exploratory approach implemented with the CROPS algorithm. Section 3.4 reviews applications of change-point detection on environmental data. it concludes on the lack of examples in the literature analysing complex data such as that described in Section 2.3 with change-point detection methods.
Chapter 4 introduces an adapted change-point detection method for concentration data that integrate left censoring in the model.

4. Change-point detection for concentration data

Contents

In this Chapter, we build a change-point detection algorithm specially adapted for concentration data. We will use this algorithm in the subsequent chapter to detect homogeneous temporal periods on which spatial statistical inferences will be possible. Several elements presented in Chapter 3 are used to build this method:

- We use a parametric change-point detection, more precisely a maximum likelihood based method as described in Section 3.1. The cost function W is defined as the negative log likelihood of a distribution Q. The choice of Q is motivated by the observation of the data, we mentioned in Chapter 2 that the distributions of concentrations were right skewed and presented long tails. Probability laws such as the Weibull are used for illustrations.
- We use the PELT search method presented in Section 3.4 to obtain optimal solution to the change-point detection problem. Several penalty values β are explored with the CROPS algorithm presented in Section 3.4. The elbow method is applied when it is necessary to estimate an optimal number of change-points.

We first describe the model integrating the censoring information in Section 4.1. However, we do not know how much the censoring can affect a parametric change-point model. We provide a study of censoring effects in Section 4.2. Furthermore, we need to devise an estimation procedure that is adapted to the observations of pesticide concentrations. We propose an original change-point detection method, involving an iterative procedure to estimate both piecewise constant parameters and parameters constant over time in the model. We devise our estimation scheme in Section 4.3. Finally, in Section 4.4, we compare our method against the *Multrank* non-parametric change-point method that can take into account the censoring in the data in its cost function (Lung-Yut-Fong et al., 2015). This method was reviewed in Chapter 3.

4.1. Generic model for censored data

4.1.1. Framework

We present here the underlying parametric model which summarizes the standard model introduced in Section 3.1. We consider $\boldsymbol{c} = c_1, \ldots, c_n$ which are realizations of independent real random variables $C_1, \ldots, C_n$. The variables C_i are recorded sequentially, and the recording times are not necessarily equidistant. Thus, the indices in C_i are only indicators of the order of occurrence in the sample and not of the observation times. We suppose that there exist K^* changes in the distribution of $\boldsymbol{c}$ happening at indices $0 = \tau_0^* < \tau_1^* < \ldots < \tau_k^* < \ldots < \tau_{K^*}^* < \tau_{K^*+1}^* = n$. Moreover, on the k-th segment, all random variables in the segment $C_{\tau_{k-1}^*+1:\tau_k^*}$ follow a distribution Q with parameters defined by the vector $\theta_k^* \in \Theta$ with $\Theta \subset \mathbb{R}^P$. We denote $\boldsymbol{\theta^*} = (\theta_k^*)_{k=0}^{K^*}$. More formally, we have that:

$$c_t \sim \sum_{k=0}^{K^*} f(.|\theta_k^*) \mathbb{1}_{\tau_k^*+1 \leq t \leq \tau_{k+1}^*},$$

f being the density function of distribution Q with respect to the Lebesgue measure on $\mathbb{R}$.

The observations are subject to censoring. We focus on left-censoring because it is adapted for modelling concentration data but similar models can be created for right censoring or a mix of both. To each c_i is associated a known censoring threshold $a_i \geq 0$. The resulting censored observations are defined by:

$$Y_i = \sup(C_i, a_i) \tag{4.1}$$

Since the C_i are independent and the a_i are known deterministic values, the Y_i are independent as well. The observations of Y_i are denoted y_i. In this model, we can write the log-likelihood of a segment $y_{\tau_k^*+1:\tau_{k+1}^*}$ as:

$$\mathcal{L}(y_{\tau_k^*+1:\tau_{k+1}^*}, \theta_k^*) = \sum_{i=\tau_k^*+1}^{\tau_{k+1}^*} \log(f_{\theta_k^*}(y_i)) = \sum_{i=\tau_k^*+1}^{\tau_{k+1}^*} \log(F(y_i|\theta_k^*))\mathbb{1}_{y_i=a_i} + \sum_{i=\tau_k^*+1}^{\tau_{k+1}^*} \log(f(y_i|\theta_k^*))\mathbb{1}_{y_i>a_i}, \tag{4.2}$$

with $f_{\theta_k^*}$ being the density function of C_i for $i \in [\tau_k^* + 1, \tau_{k+1}^*]$, $F(|\theta_k^*)$ being the cumulative distribution function (cdf) of Q.

Note that if one needs to integrate right censoring into the likelihood, one should simply replace the sup in the definition 4.1 by the inf of both quantities and cdf function F by the survival function $S(t) = 1 - F(t)$.

In practice, the values a_i correspond to the values of LOQ[1]. We could develop a double-censored model using both the LOD[2] and LOQ values. However, since the information from LOD is rarely available, we only use the LOQ.

4.1.2. Estimator choices and properties

In this section, we define the criterion to minimise and we choose the estimators of the parameters in our model.

The cost function used to evaluate segments is the negative log-likelihood calculated using the maximum likelihood estimator $\widehat{\theta}_{u:v}$ of segment $y_{u:v}$:

$$W(y_{u:v}) = -\sup_{\theta\in\Theta}\{\sum_{i=u}^{v} \log(F(y_i, \theta))\mathbb{1}_{y_i=a_i} + \sum_{i=u}^{v} \log(f(y_i, \theta))\mathbb{1}_{y_i>a_i}\}, \tag{4.3}$$

with F the cumulative distribution function (cdf) of Q.

Since we do not know the number of change point before hand, we opt for the penalized criterion defined in 3.4. For a segmentation $\mathcal{T} = \{\tau_1, ..., \tau_K\}$ of a given signal $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, the penalised cost is given by:

$$\mathcal{C}(\boldsymbol{y}, \mathcal{T}) = \sum_{i=0}^{K} W(y_{\tau_i+1:\tau_{i+1}}) + KP\beta, \tag{4.4}$$

[1]Limits of quantification (see Chapter 2.3)

[2]Limits of detection (see Chapter 2.3)

where P is the dimension of the parameters vector in the distribution Q and $\beta > 0$ the penalty weight value. Gathering (4.3) and (4.4), this resulting estimator can be expressed as:

$$(\widehat{K}, \widehat{\mathcal{T}}, \widehat{\boldsymbol{\theta}}) = \arg\min_{\mathcal{T},\boldsymbol{\theta},K} \left(-\sum_{i=0}^{|\mathcal{T}|} \left\{ \sum_{j=\tau_i+1}^{\tau_{i+1}} \log(F(y_j,\theta))\mathbb{1}_{y_j=a_j} + \sum_{j=\tau_i+1}^{\tau_{i+1}} \log(f(y_j,\theta))\mathbb{1}_{y_j>a_j} \right\} + \beta K P \right) \tag{4.5}$$

Furthermore, we assume the following on the model:

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

H2: There exists $\Delta^*_{\boldsymbol{\tau}} > 0$ such that $|\tau^\star_k - \tau^\star_{k-1}| > \Delta^*_{\boldsymbol{\tau}}$, for all $k = 1, ..., K^\star$.

H3: There exists a positive integer K_{max} such that the maximum number of segments $\frac{n}{\Delta^*_{\boldsymbol{\tau}}}$ satisfies $K_{\max} \geq \frac{n}{\Delta^*_{\boldsymbol{\tau}}}$.

H4: The penalty value is dependant of n. It can be written β_n and verifies $\beta_n \xrightarrow[n\to\infty]{} \infty$ and $\frac{\beta_n}{n} \xrightarrow[n\to\infty]{} 0$.

These are standard hypotheses for change point detection that can be found in Lavielle (1999) or He & Severini (2010). Hypothesis **H1** mainly aims at ensuring sufficient conditions for the identifiability of the model, by imposing a minimum gap between two consecutive θ's.
Hypothesis **H2** checks that each segment contains sufficient data for obtaining reliable estimates for the θ's and Hypothesis **H3** states the number of regimes is bounded from above.
Hypothesis **H4** is verified by a large range of penalties including the BIC (Yao, 1988).
We formulate an additional hypothesis **H5**, which is also the strongest one.

H5: 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 estimator, 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.
Note that in practice, for environmental data with scarce and irregular sampling, this hypothesis is hard to verify. However, we may suppose that the change-points occurring in concentration data are linked with the farming activities and the phyto-pharmaceutical uses, regardless of the sampling rate.
Under these assumptions, we know that the maximum likelihood estimator computed in Equation (4.5) is weakly consistent (Lavielle, 1999).

$$(\widehat{K}, \widehat{\mathcal{T}}, \widehat{\theta}) \xrightarrow[n\to+\infty]{\mathbb{P}} (K^*, \mathcal{T}^*, \theta^*) \tag{4.6}$$

Elements of proof are provided in Appendix A.1.

4.1.3. Estimation procedure

Estimation of parameters for a single homogeneous segment

For a given $y_{u:v}$, one should be able to properly compute $\widehat{\theta}_{u:v}$. as the estimate of the parameter of Q. It is impossible to obtain a explicit formula of $\widehat{\theta}_{u:v}$ when there are censored observations in the sample. Therefore one needs to use a numerical optimisation procedure to compute the estimate. In this thesis, the Newton-Raphson method was used to search for the zeros of the first derivative of the cost function (4.3).
For some choices of Q, checking that the cost function is strictly convex is not always easy. However, we can still show that the cost is locally convex and has a unique minimum by studying its variations. This implies a careful choice in the parameter initialization value of the Newton-Raphson method to obtain convergence of the solution $\widehat{\theta}_{u:v}$. Appendix A.2 provides experiments on the initialization of the Newton-Raphson method.

Estimation procedure for K^*, $\mathcal{T}^*$ and θ^*

The estimation procedure aims at maximising the penalised log-likelihood in Equation (4.4). The optimisation procedure implies finding both the best partitioning of the data as defined by $\widehat{K}$ and $\widehat{\mathcal{T}}$, as well as the estimates for the parameters of the distribution Q within each segment, $\widehat{\theta}$.
We proceed using the PELT algorithm (see Section 3.4). In practice, we use the following value for the cost function to evaluate a given segment $y_{u:v}$:

$$W(y_{u:v}, \widehat{\theta}_{u:v}) = \sum_{i=u}^{v} \log(F(y_i, \widehat{\theta}_{u:v}))\mathbb{1}_{y_i=a_i} + \sum_{i=u}^{v} \log(f(y_i, \widehat{\theta}_{u:v}))\mathbb{1}_{y_i>a_i}, \tag{4.7}$$

where $\widehat{\theta}_{u:v}$ is the MLE, computed numerically with the Newton-Raphson method on the segment $y_{u:v}$.

Selecting the optimal number of changes

We use the CROPS algorithm (see Section 3.4) to select the optimal number of change points in the signal. The penalty range selected investigated is derived from the BIC criterion formula (Yao, 1988). Note that this choice verifies assumption **H4**. In change point detection, the BIC penalty can be written as:

$$\beta_{BIC} = \frac{P \log(n)}{2}, \tag{4.8}$$

with P the dimension of the parameter θ and n the length of signal $\boldsymbol{y}$. The range of penalty values explored in this thesis writes under the form:

$$[\beta_{min}, \beta_{max}] = \left[\frac{\beta_{BIC}}{j}, \beta_{BIC} \times m\right], \tag{4.9}$$

with $j, m \in \mathbb{N}$. Hence, our penalty range is calibrated accordingly to Hypothesis **H4** because β_{min} and β_{max} also verify its conditions.

We choose the value of j such that the penalty $\frac{\beta_{BIC}}{j}$ is associated to what would seem to be an overfitting segmentation model. More precisely, we look for j such that associated number of change points is close to the maximal number of segments possible given the minimal segment size chosen in PELT.
Likewise, m is chosen such that the penalty $\beta_{BIC} \times m$ is associated to an underfitting segmentation model. More precisely, we look for m such that associated number of change points is close to zero.
Once we run CROPS, we obtain some penalty values $\{\beta_{min}, \dots, \beta_{max}\}$. Since each of these penalty values is associated to a segmentation cost and a number of change points, we can perform the elbow method heuristic (see Algorithm 2) on the curve of the segmentation costs plotted against the number of changes. An optimal number of break points and the associated segmentation are thus selected with this procedure.
Let us now study the effect of censoring on our procedure.

4.2. Censoring effects

The censoring can be an issue in the practical implementation of the procedure developed in Section 4.1. In particular, issues can be encountered when estimating the parameter of a fully censored segment with Newton-Raphson. Illustrating examples are provided in this section with Q set as the exponential distribution.

4.2.1. Practical problem encountered with censoring

In general, an explicit formula for the maximum likelihood estimator (MLE) is not available in presence of censored data, leading to the use of numerical methods for its computation. The Newton-Raphson method was used on each segment to compute the MLE estimate of θ.
Specifically, the case where all data in the segment $y_{u:v}$ are censored is problematic. The estimation of θ is impossible. We illustrate in Figure 4.1 with Q set as an exponential distribution of parameter $\theta = \lambda$. For a given segment $y_{u:v}$ where all observations are censored and under a censoring threshold a we have that:

$$W(y_{u:v}) = - \sup_{\lambda \in]0;+\infty[} (v-u) \log(1 - \exp(-\lambda a)) \tag{4.10}$$

This cost is always positive and decreasing to 0 when λ goes to infinity.
We made the assumption that the support Θ of the segments parameters is a compact of $\mathbb{R}^P$ in Section 4.1 . This corresponds to hypothesis **H1**. Hence, we need to decide of an upper bound for θ. We show in the next section how we proceed.

4.2.2. Introduction of an upper bound parameter on θ

We have seen in Section 4.2.1 that, for the exponential distribution example, the optimal θ tends to infinity. To solve this practical problem, we introduce an additional parameter θ_{max}

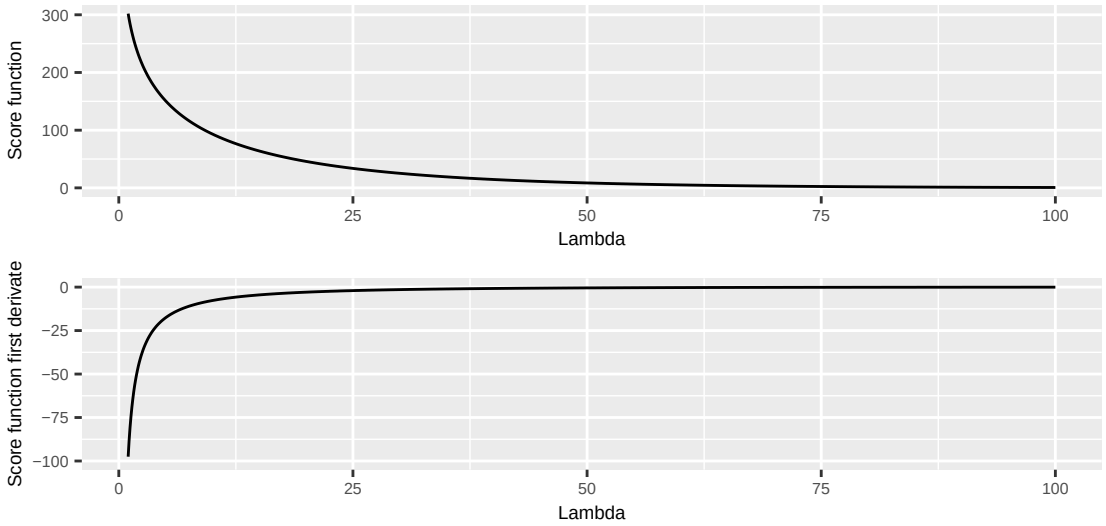

Figure 4.1: Plot of the cost function values against θ values when all observations are censored. It is represented for an exponential distribution. The sample consists in 100 values of censored observations to a threshold $a = 0.05$.

in our implementation such that θ is constrained to the interval $[0, \theta_{max}]$ which is a compact part of $\mathbb{R}$.
A new problem arises from this modelling choice. The value of θ_{max} must be chosen carefully. We must ensure that $\theta_{max} > \theta_k^*$, for all $k \in \{0, \dots, K^*\}$. If it is not the case, the identifiability problem remains. For two θ_i^* and θ_j^* greater than θ_{max}, their estimates will be set to θ_{max} and thus no segment identifiability will be possible.
In order to avoid such problems, the value of θ_{max} is set according to the worst possible censoring case. More precisely, we assume that no change-point occurred in $\boldsymbol{y}$ and that all observations are distributed according to Q with parameter θ_{max}. We set θ_{max} to the value such that:

$$F(\min(\boldsymbol{y}), \theta_{max})^n = \alpha, \tag{4.11}$$

with α being a desired percentage of censoring. In practice, we decide to set α to 95%. This corresponds to the scenario where there is a 95% chance that n observations generated from the distribution Q with parameter θ_{max} are left-censored and under the threshold $a = \min(\boldsymbol{y})$.

We provide a practical example with the exponential distribution. We simulate a signal $\boldsymbol{y}$ of size $n = 200$ that is a realization of exponential distributions of parameters $\lambda_0^* = 1$ for $y_{1:100}$ and $\lambda_1^* = 4$ for $y_{101:200}$. The censoring level is set to the median of $\boldsymbol{y}$ so that 50% of the signal is censored. We illustrate $\boldsymbol{y}$ in Figure 4.2. We choose λ_{max} by setting α to 95%. We have that $\lambda_{max} = \frac{-\log(1-\alpha^{1/n})}{\min(\boldsymbol{y})}$. In our numerical example, $\lambda_{max} = 24.68$, which is greater than λ_0 and λ_1.

Note that other ways to tackle fully censored segments are possible. The modification of the pruning rule (3.13) used in PELT can also be investigated. For example, one could decide to systematically discard all potential change-point indices $\tau \in \{u, \dots, v\}$ when evaluating a fully

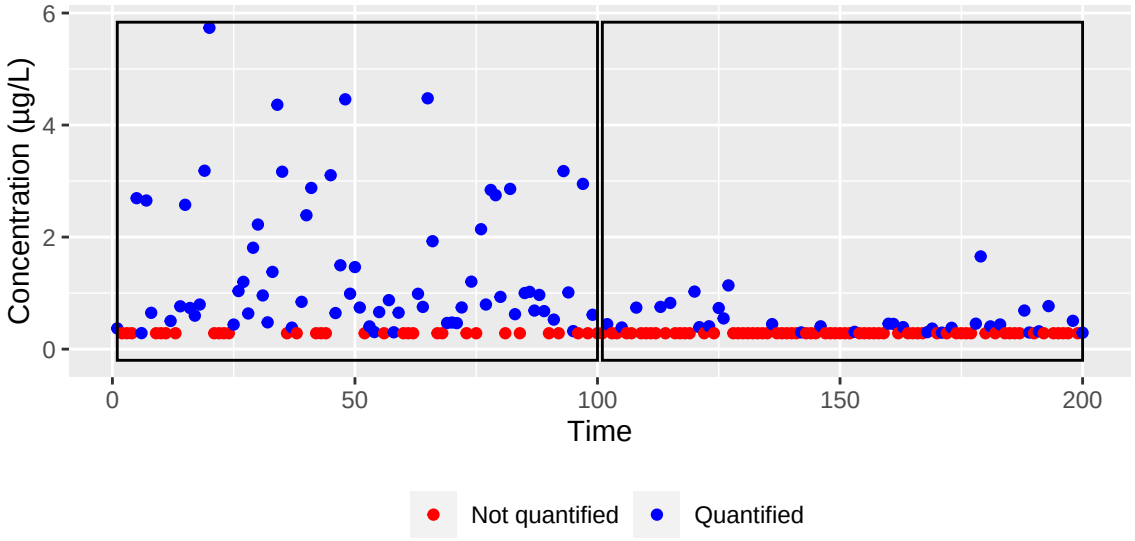

Figure 4.2: Simulated signal $\boldsymbol{y}$ with distribution Q set as the exponential law.
The two segments are drawn with black rectangles. $\theta_0^* = 1$ in the left segment, $\theta_1^* = 4$ in the right one and the censoring threshold $a = 0.28$ in both segments.

censored segment $y_{u:v}$.

4.3. Estimation with non-additive cost function: mixing fixed and changing parameter estimation

Two different cases can occur when the parameter θ is a vector of dimension $P > 1$.
In the first case, we assume that all components of the parameter θ are subject to change. This is the usual framework for change point detection as presented in Chapter 3 and Sections 4.1 and 4.2. This means that changes in different parameters of $\boldsymbol{y}$, such as the mean and variance, are detected simultaneously. In practice with censored data, it is still possible to calculate the maximum likelihood estimator of a segment using numerical methods (such as Newton-Raphson). The problem with such a modelling choice is the number of data needed in a segment to estimate all parameters. If the number of parameters to be estimated is high, the number of observations must be more important to obtain a valid statistical estimate.
In the second case, it is assumed that some dimensions of the parameter vector θ are constant over time. In this case, three scenarios are possible:

- The fixed parameters are known. This framework is then equivalent to the first case presented above. We can estimate the parameters that are changing using what is presented in Chapter 3 and Sections 4.1 and 4.2.

- The fixed parameters are unknown and we do not want to estimate them. In this case, the estimation of the changing parameters can be done with an appropriate cost function. For example, a quadratic loss function is practical for detecting changes in the mean of

a signal without estimating the variance (which is assumed to be fixed) (Fearnhead & Rigaill, 2018).

- The fixed parameters are unknown and we want to estimate them.

We are interested in the latter scenario. In the context of environmental data, the assumption that some dimensions of θ are fixed may indeed be justified if some parameter values are supposed to control the intrinsic diffusion properties of a chemical substance in the environment. Then these parameters should be invariant in time. The parameters specific to each segment could, for example, represent different intensities in the use of the substance in time.
Note that in this framework, all estimation schemes using PELT or other dynamic programming methods are not valid. These methods rely on the fact that the criterion to be optimised is additive, which is no longer the case. There are other methods for estimating signal changes that do not require this property, such as MCMC algorithms (Lavielle & Lebarbier, 2001) or genetic algorithms, but they are outside the scope of this thesis.
This section develop our own estimation strategy.

4.3.1. Estimators of a segmentation with some fixed parameters

We use the following notations in this section:

- $\boldsymbol{\theta}_{.,m} = (\theta_{0,m}, \dots, \theta_{K,m})$ is the m-th dimension of the parameter vector of each segment k.
- $\boldsymbol{\theta}_{k,.} = (\theta_{k,1}, \dots, \theta_{k,P})$ is the parameter vector of the k-th segment.
- $\theta_{k,m}$ is the m-th dimension of the parameter vector of the k-th segment.

We are interested in models where changes occur only in some dimensions of $\boldsymbol{\theta}^*$. We denote $\mathcal{M} \subset \{1, \dots, P\}$ the set of indices of dimensions where the changes occur and $\overline{\mathcal{M}}$ the complementary set. We introduce the notations $(\boldsymbol{\theta}^*_{.,m})_{m\in\mathcal{M}} := \boldsymbol{\theta}^*_{\mathcal{M}}$ and $(\boldsymbol{\theta}^*_{.,m})_{m\in\overline{\mathcal{M}}} := \boldsymbol{\theta}^*_{\overline{\mathcal{M}}}$
Therefore, the parameters $\boldsymbol{\theta}^*_{\overline{\mathcal{M}}}$ are reduced to a vector of fixed parameters throughout the signal $\boldsymbol{y}$. In this setting, the estimation procedure differs from Equation (4.5) and can be expressed as:

$$(\widehat{K}, \widehat{\mathcal{T}}, \widehat{\theta}_{\mathcal{M}}, \widehat{\theta}_{\overline{\mathcal{M}}}) = \arg \min_{K,\mathcal{T},\boldsymbol{\theta}_{\mathcal{M}},\theta_{\overline{\mathcal{M}}}} \left[\left\{ - \sum_{i=0}^{|\mathcal{T}|} \left(\sum_{j=\tau_i+1}^{\tau_{i+1}} \log(F(y_j, \theta)) \mathbb{1}_{y_j = a_j} + \sum_{j=\tau_i+1}^{\tau_{i+1}} \log(f(y_j, \theta)) \mathbb{1}_{y_j > a_j} \right) \right\} + \beta K P \right] \tag{4.12}$$

Unfortunately the criterion in Equation (4.12) is not additive as constant parameters introduce dependencies between the costs of each time segment. We cannot leverage the additivity property to obtain fast algorithms based on dynamic programming. We propose therefore to use an alternating optimisation procedure described in the following section.
From Equation (4.12), we can now design a two-step iterative estimation strategy.

4.3.2. Estimation procedure

We describe the practical implementation of the estimators defined in Equation (4.12). We choose to proceed with two nested steps: the inner step uses an alternating optimization scheme to solve problem (4.12) for a fixed value of β while the outer step explores a range of values for β.

Inner step

For a fixed penalty value β in a range $[\beta_{min}, \beta_{max}]$ and an initial $\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{M}}}$, we iterate the following two steps until convergence:

1. Run the PELT algorithm with penalty β and fixed $\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{M}}}$ to estimate $\widehat{\mathcal{T}}$ and $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}$.
2. Compute the MLE $\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{M}}}$ with fixed $\widehat{\mathcal{T}}$ and $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}$ with any optimization method that handles censored data. We use the `R` package developed in Delignette-Muller & Dutang (2015) in our procedure.

The initialization is an important part of this procedure. We would like the initial estimate of the fixed parameters to be close to $\boldsymbol{\theta}^*_{\overline{\mathcal{M}}}$ to ensure the convergence of the procedure. In order to do so, we initialise $\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{M}}}$ assuming no change-point occurred in $\boldsymbol{y}$. More precisely, in this case $\boldsymbol{\theta}^*$ is a vector and we use the MLE estimator:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\theta} \left\{ -\left(\sum_{j=1}^{n} \log(F(y_j, \theta)) \mathbb{1}_{y_j = a_j} + \sum_{j=1}^{n} \log(f(y_j, \theta)) \mathbb{1}_{y_j > a_j} \right) \right\}, \tag{4.13}$$

which implies using iterative methods again as stated in Cohen (1965).
The initial value of the vector $\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{M}}}$ is initialized with the corresponding coordinates in $\widehat{\boldsymbol{\theta}}$. Note that the initialization does not depend on the penalty value β.

Outer step

This step aims at selecting penalty values β in a given range $[\beta_{min}, \beta_{max}]$.
A first simple approach is to define an evenly spaced penalty grid $(\beta_{min}, \ldots, \beta_q, \ldots, \beta_{max})$ where $\beta_{min} < \cdots < \beta_q < \cdots < \beta_{max}$. The inner step is performed on each of these β_q. Eventually, the optimal penalty value, hence the corresponding parameters, is selected using an elbow rule heuristic as proposed in Section 3.3.
Another option is to use the CROPS algorithm to explore the penalty range since this algorithm only requires the penalty term to be linear and number of breaks to be non increasing with respect to the penalty weight, without any other specific assumption on the cost function. In particular, it does not require the cost function to be additive.

We verify in Appendix A.3 that the theoretical conditions needed for PELT to provide an optimal solution are fulfilled in the inner step. The next section is devoted to the evaluation of the efficiency of our method through numerical simulations. This optimisation method will be used later in Chapter 5.

4.4. Simulation study

In this section we conduct experiments with simulated data. For all simulated data, the distribution Q is set as a Weibull law with the scale parameter λ and the shape parameter σ. The data are simulated so that λ is the parameter whose value changes, while σ is constant across the simulated signals.
First, we want to calibrate the minimum segment size parameter in the PELT search method. We recall that this method is used for the segmentation step (inner step 4.3.2) of our algorithm. We define a simple framework in these experiments by assuming that the value of σ is known. Comparisons are made with the *MultRank* non-parametric method (Lung-Yut-Fong et al., 2015) to establish a satisfactory value for the minimum segment size (see Section 4.4.1).
We then compare the performance of parametric change point detection with the *MultRank* method once again. The σ should also be known in this framework. This time, these experiments are performed to confirm that calibration of the minimum segment length value in 4.4.1 results in correct change point detection. More complex signals are simulated, e.g. signals with multiple change points (see Section 4.4.2).
Finally in section 4.4.3, we discuss the results of the estimation procedure developed in Section 4.3.2. In this case, the shape parameter σ is unknown and must also be estimated with the positions (and the number) of the change points and the scale parameters of each segments.

4.4.1. Calibration of the minimum segment size

In this section, we use the Weibull distribution for the distribution Q. The parameters of the simulated signals are:

- n the signal size.
- K^* the number of changes in the scale parameter values.
- $(\tau_k^*)_{k=1}^{K^*}$ the position of these changes.
- $(\lambda_k^*)_{k=0}^{K^*}$ the scale parameters of each segment.
- σ^* the known shape parameter.
- α the censoring rate of the signal.

Note that α is a global *a posteriori* censoring level. For a given signal $\boldsymbol{y}$ and α censoring level, the resulting censoring threshold is the empirical α-quantile of $\boldsymbol{y}$.

In section 2.2, the PELT Algorithm 3 introduces a minimal segment length n_{min}. We conduct some simulation tests to calibrate this parameter. We need to identify n_{min} so that the cost function defined in (4.3) has sufficient data to detect a change between two segments with different parameters. This can be viewed as a classification task. We want to know if our method is able to correctly classify signals that have a change point or not.

It can be noted that the PELT detection ability and the likelihood ratio statistic (LR) are linked when dealing with a single change point. Let $\boldsymbol{y}$ be a signal of size n. The LR statistic is used to test the hypothesis::

$\mathcal{H}_0$: there is no change point in $\boldsymbol{y}$ and $y_1, \ldots, y_n$ was generated by the law Q_0.

$\mathcal{H}_1$: there exists an index t such that $y_1, \ldots, y_t$ was generated by the law Q_1 and $y_{t+1}, \ldots, y_n$ by the law Q_2.

Using the notations from Equation (4.3), the LR statistics writes as:

$$\Lambda = 2(W(y_{1:n}) - (W(y_{1:\widehat{t}}) + W(y_{\widehat{t}+1:n}))), \tag{4.14}$$

where $\widehat{t}$ is the estimated change point location under hypothesis $\mathcal{H}_1$. $\widehat{t}$ can be computed with the optimal partition Algorithm 1.
Under $\mathcal{H}_0$, we can expect low values of Λ. On the contrary, high values of Λ will lead to question the validity of $\mathcal{H}_0$. Looking at the inequality $\Lambda \geq B$ with $B > 0$, we can finally write that:

$$\Lambda \geq B \iff W(y_{1:n}) \geq W(y_{1:\widehat{t}}) + W(y_{\widehat{t}+1:n}) + \frac{B}{2} \tag{4.15}$$

Equation (4.15) shows that comparing the likelihood ratio Λ to a threshold B is equivalent to comparing the penalized cost of a single change-point model for $\boldsymbol{y}$ with penalty value $\frac{B}{2}$ to the cost of the whole signal without change-point.
We can then achieve a diagnosis of the classification ability of the PELT method building ROC curves of likelihood ratio statistics (Fawcett, 2006) e.g. studying the variation of false positive and true positive rates when the decision threshold B varies. Summarising the ROC curve with the Area Under Curve criteria gives a general idea of the classifying performance of the likelihood ratio statistic (thus of the PELT method).
A comparison of our method is made with the *MultRank* method that is based on non parametric inference that we presented in Section 3.1. More precisely, this method relies on the statistic S_n defined in Equation (3.7). The decision in this statistical test is also made by comparing this statistic to a threshold B. Thus, we can also compute a ROC curve from several S_n statistics.
We want to assess the detection ability of the PELT method in function of its minimum segment size. The simulation protocol is the following:

1. For a given value of n_{min}, we simulate M signals of size n. $\frac{M}{2}$ signals are simulated according to hypotheses $\mathcal{H}_0$ and $\frac{M}{2}$ according to $\mathcal{H}_1$.

2. For each of these signals, we compute the LR statistic Λ and the S_n statistic of the *Multrank* method. For the parametric case, we compute the optimal segmentation with a single change point using the optimal partition Algorithm 1 with the minimum segment size constraint n_{min}. This constraint is also implemented on the calculation of the S_n statistic.

3. From the M LR statistics and the M S_n statistics, we construct the ROC curves for both methods and we summarise it calculating its AUC.

We test several configurations with different values of $n_{min} = (5, 10, 25, 50, 75)$ and different censoring levels $\alpha = (5\%, 25\%, 50\%, 75\%, 95\%)$. For each configuration, we simulate $M = 1000$ signals of size $n = 200$.
The parameters used to simulate under hypothesis $\mathcal{H}_0$ are: $\sigma = 0.5$, $\theta_0 = 1$. Under hypothesis $\mathcal{H}_1$, we set the parameters as: $\sigma = 0.5$, $\lambda_1^* = 1$, $\lambda_2^* = 3$, $t = 100$.

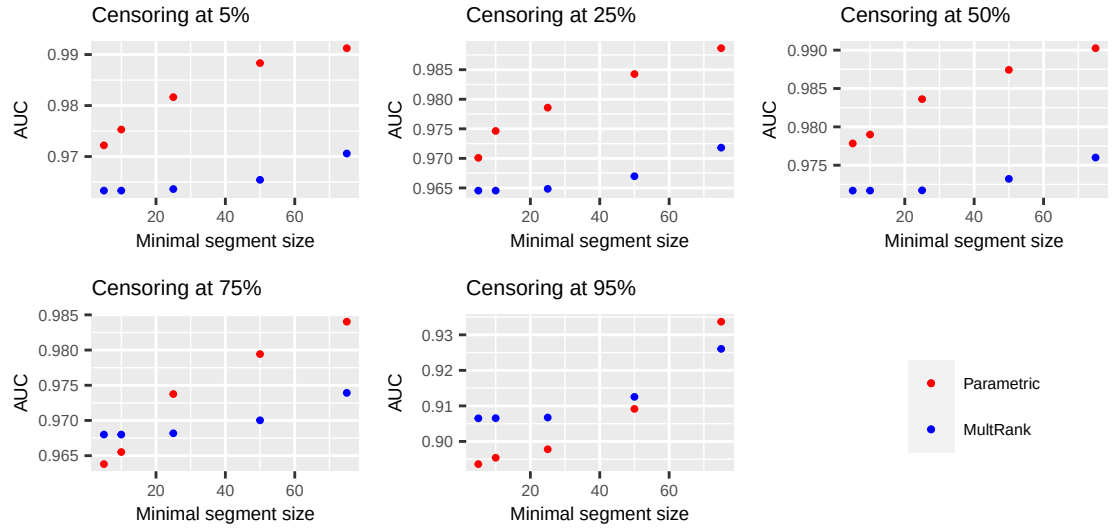

Figure 4.3: Choice of the minimal segment length: simulation results. Our method performance is illustrated with the red dots, the *Multrank* method is drawn in blue.

From the results of Figure 4.3, three comments can be made:

- Both methods are efficient in their abilities to detect change-points in signal. The AUC are above 0.9 (except for the parametric method in a highly censored configuration with a low value minimal segment size). That make them both good classifiers.
- The performance of the parametric method increases with the minimum segment length.
- The more the censoring level increases, the more data is needed in the minimal segment length of the parametric method to outclass the non parametric method.

We summarize our results in the two following points. Firstly, the minimum segment size depends on the censoring level. Secondly, the parametric method outperforms *Multrank* for large enough minimum segment size even up to 75% of censoring level.
It should be noted that we chose an easy simulation scenario on purpose. The parameters on each side of the change point (in the signals simulated under $\mathcal{H}_1$) do not have close values, the position of the change point is right in the middle of the signals and the data is truly distributed according to a Weibull distribution. This can explain the better classification using the likelihood ratio statistic.
However, we can also say that if the Weibull model seems to be a good model for a real dataset, we can expect good performances on the detection ability of the parametric method.

4.4.2. Testing the precision of the detection method with known σ

We want to compare the performance of the change-point detection with the Multrank method developed in Lung-Yut-Fong et al. (2015) since both methods are adapted to censored data. We examine the capacity to estimate the correct number of breaks in a signal and the precision of the change-point position. This section is illustrated using the Weibull distribution.
The experimental framework is as follows:

1. we simulate $M = 100$ 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 censoring 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 $\tau_1^\star = 80$, $\tau_2^\star = 160$, $\tau_3^\star = 240$ and $\tau_4^\star = 320$ and take the values $\boldsymbol{\lambda^\star} = (\lambda_1^\star = 1, \lambda_2^\star = 4, \lambda_3^\star = 0.5, \lambda_4^\star = 5, \lambda_5^\star = 1)$.

2. For each of the M samples, we apply 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 $(\hat{\tau}_{k,param})_{k=1}^{\hat{K}_{param}}$ (respectively $(\hat{\tau}_{k,multrank})_{k=1}^{\hat{K}_{multrank}}$). We set the minimum segment size $n_{min} = 25$. We have seen in the previous section that this is still a minimum segment size where the parametric method outperforms the *MultRank* (except in highly censored signals).

3. For both methods, we count the number of samples among the M for which the correct number of breaks has been estimated (e.g. $\hat{K}_{param} = K^\star$). Also, we make a histogram of the change-point positions of the samples for which $K^\star$ is estimated correctly.

Since K is not known, we proceed as follows for each method to estimate it:

- For the parametric method: we use the 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 Haynes et al. (2017). The choice of β_{min} and β_{max} is inspired from linear penalties like the BIC criterion Yao (1988). Note that when using the BIC penalty in change point detection, the penalty term written in Section 4.1 becomes : $\beta_n = \frac{P}{2}\log(n) = \frac{1}{2}\log(n)$, where P 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 using the optimal partition search method presented in Algorithm 1 for K_{max} change-points. For each of these segmentations, we can compute the cost of the segmentations using the cost function in (3.9). As in the parametric method, we represent the costs as a function of the number of change points, 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 4.1 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.

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

Table 4.1: Number of correct estimations of K over $M = 100$ samples for both methods for different $\alpha\%$ censoring rates.

4.4.3. Testing the estimation procedure with an unknown σ

We test the procedure developed in Section 4.3.2 on two different simulated signals $\boldsymbol{y}$ and $\boldsymbol{z}$. We set Q as the Weibull distribution with λ denoting the scale parameter and σ the shape parameter. We suppose that the changes occur in λ and that σ is fixed throughout the signal. The two simulated signals with their respective parameter information are illustrated in Figures 4.5 and 4.6.
$\boldsymbol{y}$ is $n = 500$ sized signal. The change point are located at position 120, 185, 310, 365. The shape parameter is set to $\sigma = 0.33$. The scale parameters on each segment are $(\lambda_k^*)_{k=0}^5 = (2, 1, 3, 1, 2)$. Every value less than the median of $\boldsymbol{y}$ was censored corresponding to the censoring threshold $a = 0.67$.
$\boldsymbol{z}$ is a $n = 1000$ sized signal. The change point are located at position 120, 185, 310, 365, 495, 580, 700 and 850. The shape parameter is set to $\sigma = 0.33$. The scale parameters on each segment are $(\lambda_k^*)_{k=0}^8 = (1/50, 1/10, 1/100, 1/5, 1/100, 1/20, 1/70, 1, 1/50)$. Every value less than the median of $\boldsymbol{z}$ was censored corresponding to the censoring threshold $a = 0.01$.
Both of outer step strategies presented in Section 4.3.2 are implemented to explore the range penalty $[\beta_{min}, \beta_{max}]$:

- We use a penalty range with 10 evenly spaced values of β.
- We explore the penalty range using CROPS algorithm rules to uncover new values of β.

The values of β_{min} and β_{max} are set according to the signal lengths. We choose these values so that β_{min} corresponds to an overfitting model and β_{max} an underfitting one. More precisely, the values are set to $\beta_{min} = \frac{\log(n)}{10}$ and $\beta_{max} = \log(n)$ for $\boldsymbol{y}$; $\beta_{min} = \frac{\log(n)}{4}$ and $\beta_{max} = 2\log(n)$ for $\boldsymbol{z}$.
The results for the penalty values tested for both signals are illustrated in Figures 4.7 and 4.8. These results show that:

- As expected, exploring the penalty range using CROPS is faster than using the penalty grid. Moreover, exploring the penalty range with a regular step can lead to a sub optimal exploration of the penalty values. In Figures 4.8 (top left) and 4.7 (top left), we can see that the estimation procedure was run for several penalty values resulting in the same segmentation. CROPS uncovers segmentation results that can be missed with a regular grid.

- When the correct model is present among the uncovered segmentations, the associated estimated parameters are accurate (change point locations, segment parameters and common shape parameter). We can conclude that the estimation procedure is correct provided the correct model is selected (see Figures 4.9 and 4.10). The estimated values of σ remain close to its true value even when the incorrect number of change-point is selected. Using the results of the CROPS procedure, the estimated σ ranged between 0.334 and 0.342 for signal $\boldsymbol{y}$; they ranged between 0.308 and 0.348 for $\boldsymbol{z}$.

- The selection of the segmentation model is made with the elbow heuristic (Figures 4.7 top and bottom right and 4.8 top and bottom right). We can see that it did not select the correct model except when using the penalty grid for signal $\boldsymbol{y}$. We can argue that the correct model was selected by chance in this case because of the step used in the grid. Since we missed some segmentation results, the elbow was more apparent. However, we can see that the elbow method selected a model that was very far from the true setting using the grid in the case of $\boldsymbol{z}$. The model selected with the CROPS results missed some change points both in $\boldsymbol{y}$ and $\boldsymbol{z}$. The change points in positions 467 and 580 were not detected in $\boldsymbol{z}$. The change point in position 133 was not detected for $\boldsymbol{y}$. The λ parameters surrounding these change-points in each signal have very close values which can explain the fact they were not detected.

We can conclude that the estimation procedure is working accurately on simulated data. However, the model selection is a complex task only partially fulfilled by the elbow heuristic. Note that other heuristics can be used to select the optimal model.
In practice, this is not a major issue since we aim at providing several models to the experts. We use the estimation procedure described in Section 4.3 in the subsequent chapter on a real concentration data. For presentation purposes, we will use the elbow heuristic to select a segmentation on these data as well but we keep in mind that the resulting segmentation may not be the optimal one and that neighbouring segmentation should be also explored. In Chapter 6, we overcome this selection problem by presenting all segmentation models found in an interactive application.

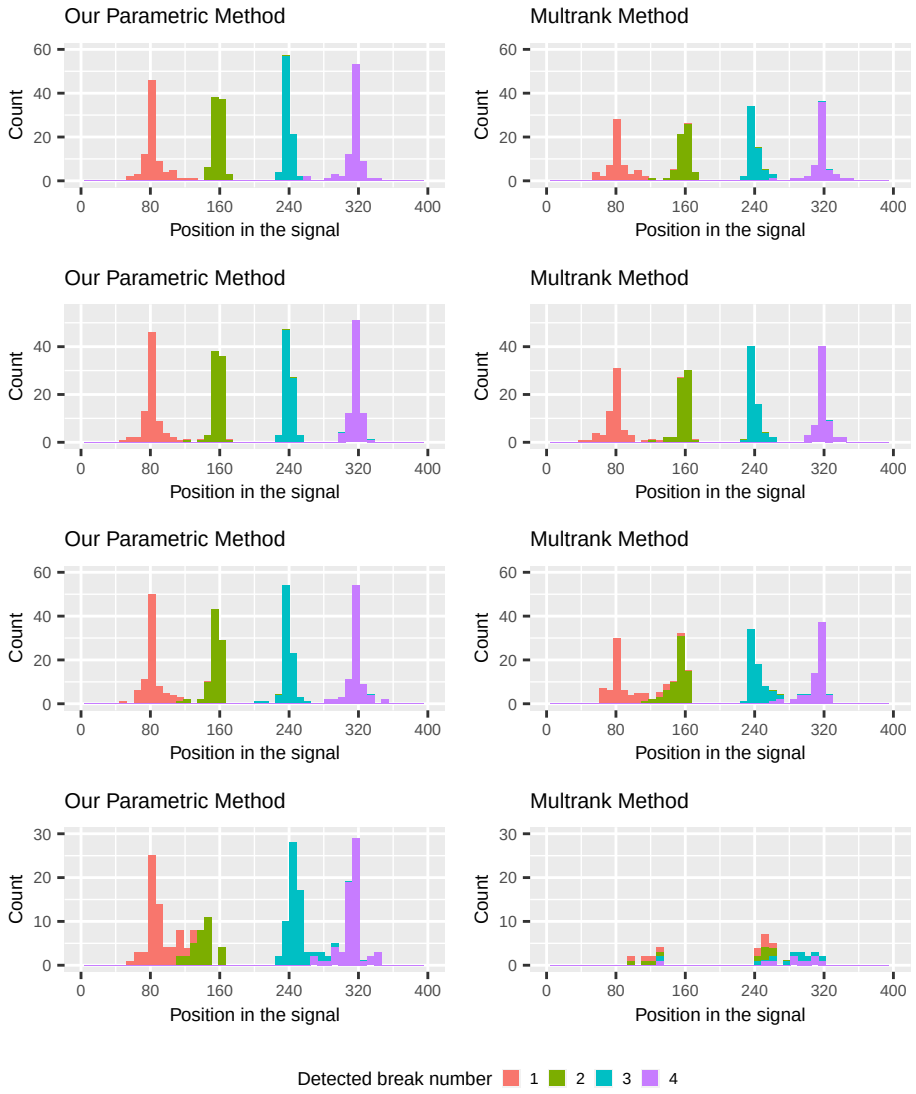

Figure 4.4: Precision of the estimated change-points for both methods. Each row corresponds to a different level of censoring α increasing from 25% (first row) to 95% (fourth row).

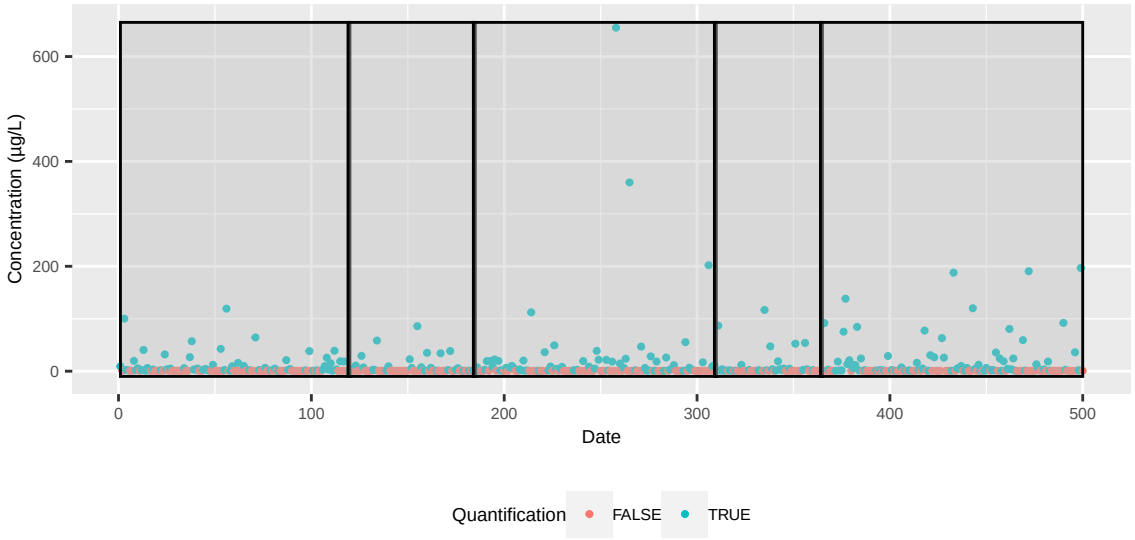

Figure 4.5: Simulated signal $\boldsymbol{y}$ of size $= n = 500$. The change point are located at position $120, 185, 310, 365$. The segments are illustrated in black. The shape parameter is set to $\sigma = 0.33$. The scale parameters on each segment are $(\lambda_k^*)_{k=0}^5 = (2, 1, 3, 1, 2)$. Every value less than the median of $\boldsymbol{y}$ was censored corresponding to the censoring threshold $a = 0.67$.

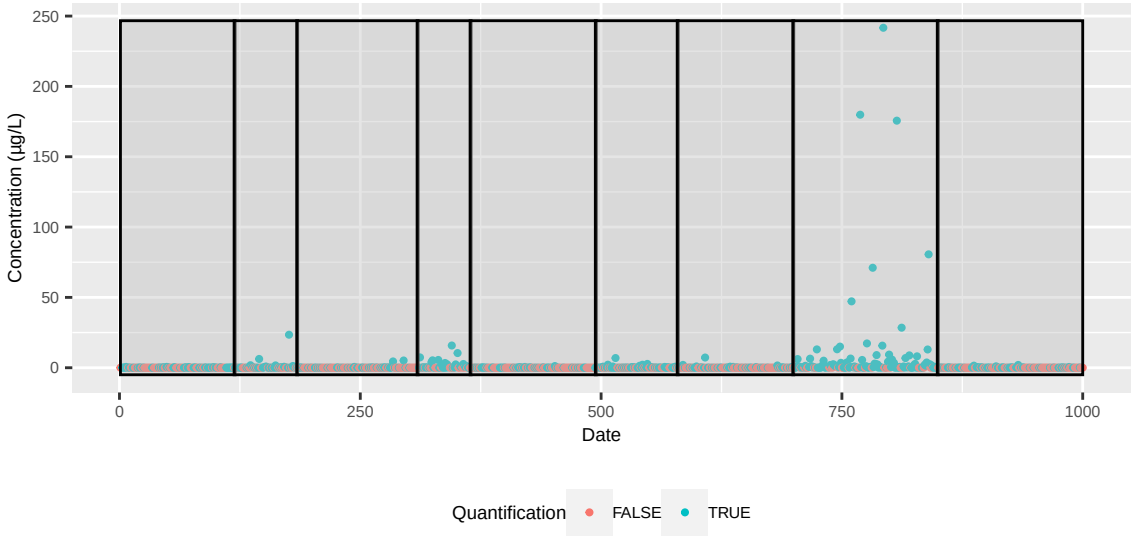

Figure 4.6: Simulated signal $\boldsymbol{z}$ of size $n = 1000$. The change point are located at position $120, 185, 310, 365, 495, 580, 700, 850$. The segments are illustrated in black. The shape parameter is set to $\sigma = 0.33$. The scale parameters on each segment are $(\lambda_k^*)_{k=0}^8 = (1/50, 1/10, 1/100, 1/5, 1/100, 1/20, 1/70, 1, 1/50)$. Every value less than the median of $\boldsymbol{z}$ was censored corresponding to the censoring threshold $a = 0.01$.

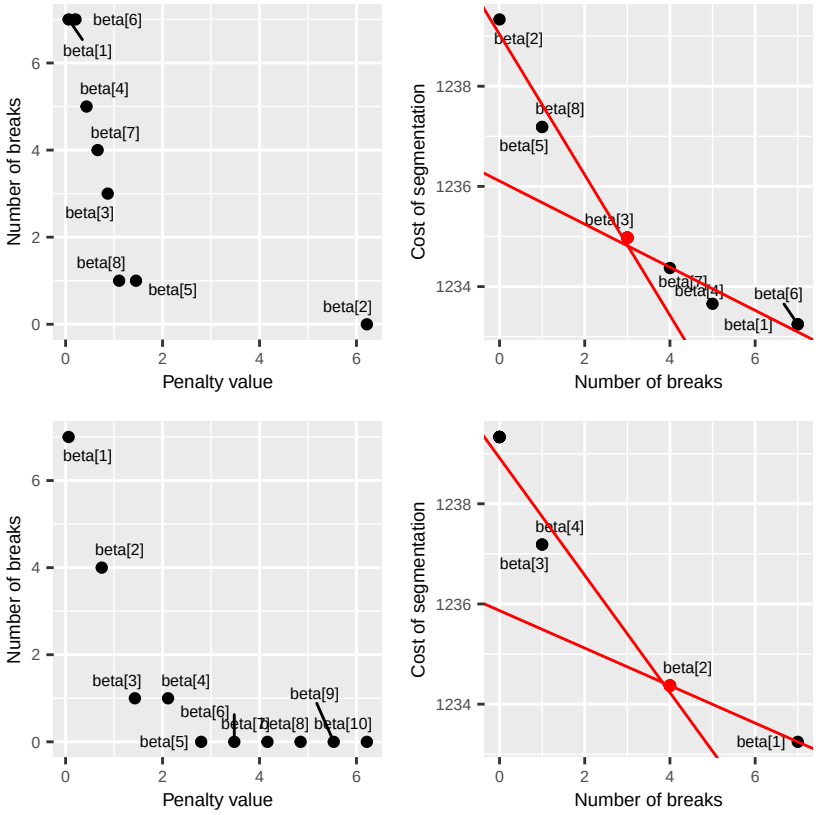

Figure 4.7: Simulation results for signal $\boldsymbol{y}$.
Top left figure: plot of the penalty values uncovered by the CROPS procedure. The number of change-points are plotted against their associated penalty values. The annotations give the order in which the penalty were uncovered.
Top right figure: plot of the cost of the resulting segmentations uncovered by the CROPS procedure. The cost of the segmentations are plotted against their associated number of breaks. The red lines represent the two part linear model chosen with the elbow heuristic. The annotations are a reminder of the penalty values in the top left figure to which the segmentations are associated to.
Bottom left figure: plot of the penalty values uncovered using the grid of penalty values. The number of change-points are plotted against their associated penalty values. The annotations give the order in which the penalty were uncovered.
Bottom right figure: plot of the cost of the resulting segmentations uncovered by the penalty grid. The cost of the segmentations are plotted against their associated number of breaks. The red lines represent the two part linear model chosen with the elbow heuristic. The annotations are a reminder of the penalty values in the bottom left figure to which the segmentations are associated to.

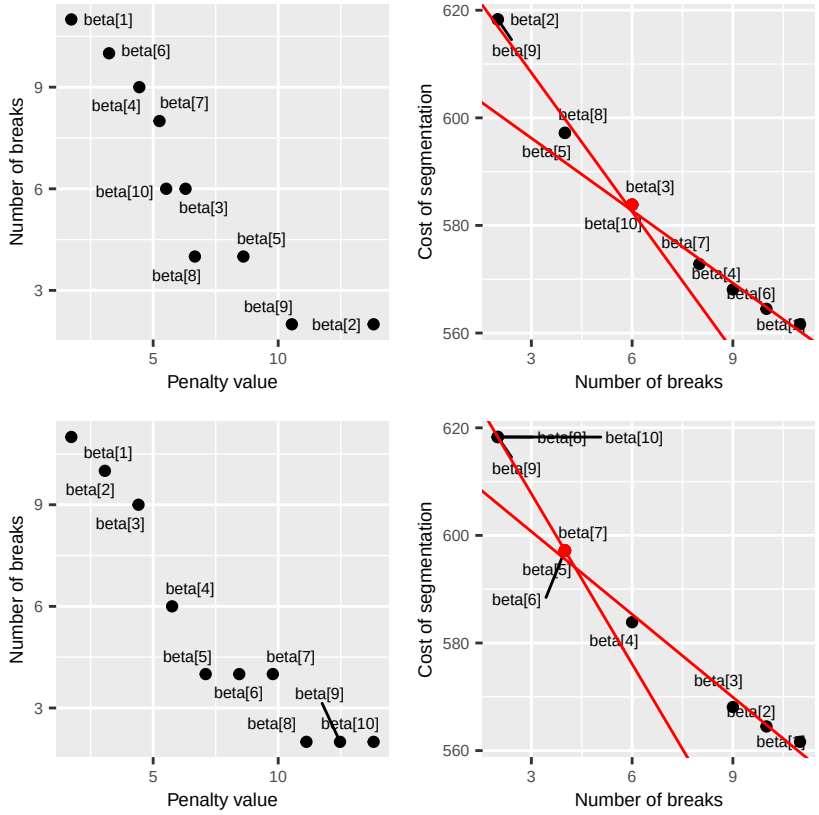

Figure 4.8: Simulation results for signal $\boldsymbol{z}$.
Top left figure: plot of the penalty values uncovered by the CROPS procedure. The number of change-points are plotted against their associated penalty values. The annotations give the order in which the penalty were uncovered.
Top right figure: plot of the cost of the resulting segmentations uncovered by the CROPS procedure. The cost of the segmentations are plotted against their associated number of breaks. The red lines represent the two part linear model chosen with the elbow heuristic. The annotations are a reminder of the penalty values in the top left figure to which the segmentations are associated to.
Bottom left figure: plot of the penalty values uncovered using the grid of penalty values. The number of change-points are plotted against their associated penalty values. The annotations give the order in which the penalty were uncovered.
Bottom right figure: plot of the cost of the resulting segmentations uncovered by the penalty grid. The cost of the segmentations are plotted against their associated number of breaks. The red lines represent the two part linear model chosen with the elbow heuristic. The annotations are a reminder of the penalty values in the bottom left figure to which the segmentations are associated to.

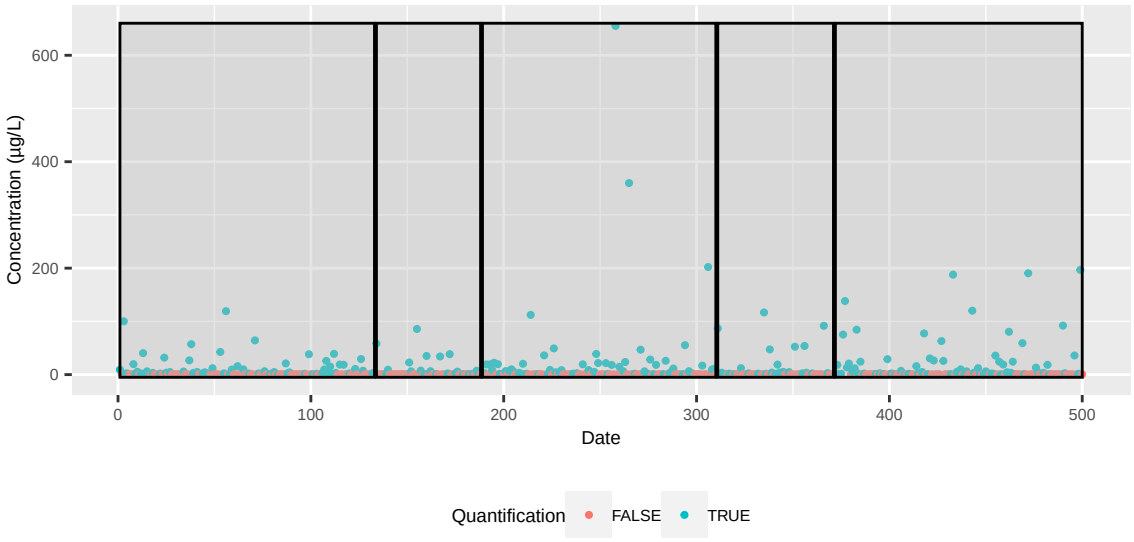

Figure 4.9: Segmentation results with correct number of change-points for signal $\boldsymbol{y}$. The estimated change-points are located at positions $133, 188, 310, 371$. The (approximated) estimated scale parameter values are $(\widehat{\lambda}_k)_{k=0}^4 = (1.62, 1.05, 2.69, 1.49, 2.99)$. The estimated shape parameter value is $\widehat{\sigma} = 0.34$.

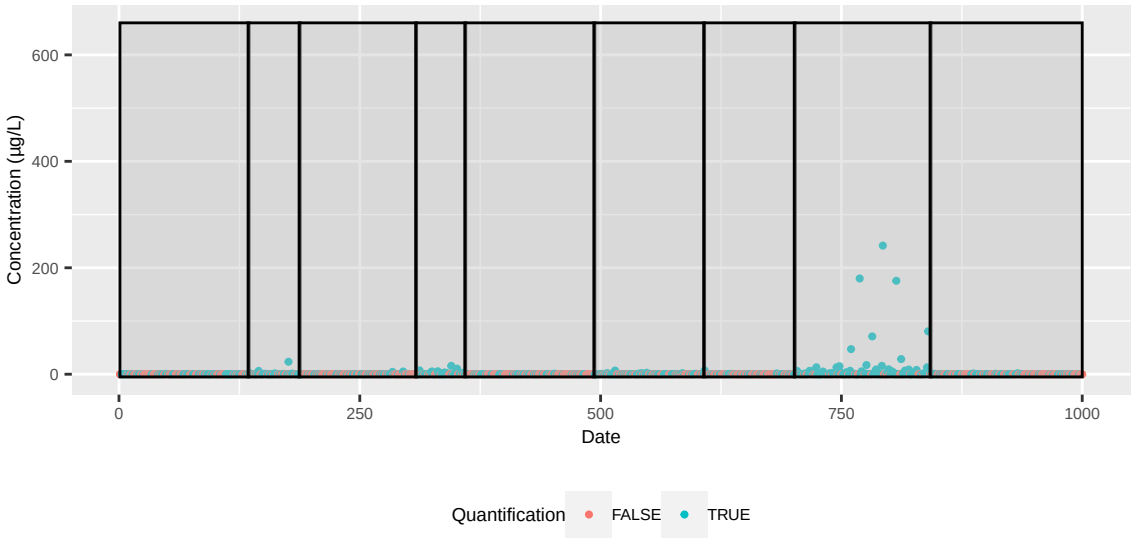

Figure 4.10: Segmentation results with correct number of change-points for signal $\boldsymbol{z}$. The estimated change-points are located at positions $134, 187, 308, 359, 493, 607, 701, 842$. The (approximated) estimated scale parameter values are $(\widehat{\lambda}_k)_{k=0}^8 = (\frac{1}{55}, \frac{1}{10}, \frac{1}{78}, \frac{1}{4}, \frac{1}{92}, \frac{1}{19}, \frac{1}{74}, 1, \frac{1}{56})$. The estimated shape parameter value is $\widehat{\sigma} = 0.34$.

4.5. Chapter summary

This Chapter designs and tests an adapted change-point detection method for concentration data. The censoring is handled by the cost function choice in Section 4.2. More precisely, using a parametric approach as presented in Chapter 3, the likelihood is adapted to distinguish cases were a measure is censored or not. The censoring becomes critical for computing the maximum likelihood estimate of a segment when all observations are censored. This could cause issues in the identifiability of segments parameters which would compromise the detection method capacity. One way to circumvent this problem is to introduce a new regularization parameter in the detection that is a maximum value for the parameter value. The estimation strategy is also discussed in this Section 4.3. An estimation scheme where some dimensions of the parameter vector are fixed in time and other can vary across segments is devised. Experiments of Section 4.4 ensure that, if enough data is available to evaluate a segment, the parametric method performs decently and can outclass a non parametric method that can also take censored data into account.
Chapter 5 combines the results of temporal change-point detection using the parametric method developed in this Chapter with statistical methods to deal with the spatial heterogeneity. The obtention of homogeneous temporal sub signal in the concentrations values provides the temporal context in which the spatial analysis is conducted. In particular, in this homogeneous temporal setting, it is interesting to look for geographical areas that had concentration values that differ from others. Constructing these areas and comparing them is the main topic of Chapter 5. Chapter 6 is the presentation of the results of Chapter 5 and 4 applied to the concentration data of substance. The results of all methods are all gathered using an interactive presentation tool.

5. Spatio-temporal analysis of concentration data

Contents

This chapter tackles 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.
While the literature on change-point detection is abundant (see Chapter 3), applications to spatial data are somewhat limited. An early example of such method can be found in Majumdar et al. (2005) while recent advances in a setting close to ours are presented in J. Chen et al. (2020). 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). This is why we propose an approach that decouples time and space in our analysis before mixing both aspects to build an anomaly criterion.
Our approach for the spatio-temporal concentration data can be described as follows.
We aggregate the complete concentration series to form the daily maximum concentration value series. We assume this newly created series to be strictly stationary, conditionally to a (possibly unknown) number of change-points and their associated locations, and the characteristics of the probability distribution within each temporal segment.
In order to model the geographical aspect, the monitoring stations are represented by a graph that integrates environmental information in its edge weights. Indeed, as geological, terrain and climatic characteristics of a given area can influence the dispersion of a chemical substance and also impact its potential use in the case of e.g. a pesticide. Therefore 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. J. Chen et al. (2020). 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.
Once this modelling for the time and space aspects is set, the methodology we propose in this Chapter takes place in three steps:

1. The change-point detection developed in Chapter 4 is used for modelling temporal heterogeneity. It produces temporal segments in which the aggregated series of daily maximal pesticide concentrations are assumed to follow a stationary distribution.

2. In parallel, clustering is used for modelling the expected spatial homogeneity while integrating geographical constraints such as river networks.

3. Conditionally to the temporal segment detected by the change-point procedure, and to the spatial cluster detected by the clustering procedure, we propose to analyse the data in order to identify contextual anomalies.

This method is inspired from previous works (Laroche et al., 2022a) but differs on a specific detail: steps 1 and 2 are independent. The clustering and change point detection can be run simultaneously.
We illustrate the whole procedure with a real data set for Prosulfocarb, a substance commonly used in France, whose concentration is monitored in surface waters. Prosulfocarb is a herbicide used to protect wheat and barley from weeds. However, its presence in water bodies can be toxic to aquatic fauna.

The rest of the chapter is organised as follows: in Section 5.1, the 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, also denoted as the fine-grain level, consists of a network made of monitoring stations, where each station is associated to an irregularly sampled univariate time-series representing recorded measures of pesticide. The second level, also denoted as the coarse-grain level, summarises the maximum recorded values of concentration 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 $V = (v_i)_{1 \leq i \leq I}$ 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 Shuman et al. (2013), 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 R non connected components that we will denote $(\mathcal{K}_1, ..., \mathcal{K}_R)$.

5.1.2. Construction of the aggregated series of maximum concentration

Each station v_i is supposed to be associated to a time series $(y_{ij}, t_{ij})_{1\leq j\leq n_i}$, where n_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 a_{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

$$\mathcal{D} = \Big((y_{ij}, t_{ij}, a_{ij})_{1\leq j\leq n_i} \Big)_{1\leq i\leq I}.$$

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 n_i (see Figure 2.2 of Chapter 2). 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 $(\eta_l)_{1\leq l\leq n}$, at which at least one observation is available in the data collection. We denote $t_{ij} \in \eta_l$ the fact that the observation time t_{ij} is compatible with η_l at the specified resolution, e.g. that the observation y_{ij} was made during the day η_l.
Third, once $(\eta_l)_{1\leq l\leq n}$ has been computed, one may introduce a coarse-grain, global series, summarising the maximum values recorded within the temporal resolution with

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

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

$$\overline{q}_l = \max \{ a_{ij} \mid t_{ij} \in \eta_l \}. \tag{5.2}$$

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

$$\overline{\mathcal{D}} = (\overline{y}_l, \eta_l, \overline{q}_l)_{1\leq l\leq n}. \tag{5.3}$$

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 $K^* > 0$ change-points producing $K^* + 1$ stationary intervals defined by

$$0 = \tau_0^* < \tau_1^* < \ldots < \tau_{K^*}^* < \tau_{K^*+1}^* = n;$$

- the observations $(\overline{y}_l)_{1 \leq l \leq n}$ are realisations of n independent random variables $(\overline{Y}_l)_{1 \leq l \leq n}$;
- when $l \in [\tau_{k-1}^* + 1, \tau_k^*]$, $\overline{Y}_l$ is distributed according to a distribution Q with parameter vector $\boldsymbol{\theta}_k^*$. Notice that, just as discussed in Section 4.3, some dimensions of the parameter vector $\boldsymbol{\theta}_k^*$ are supposed to be fixed throughout segments.

Several comments need to be made at this point. First, note that the model only takes into account the concentrations $\overline{y}_l$ but not the instants and the quantification limits, which are assumed to be deterministic quantities. The second remark is that the estimators of $(\tau_k^*)_{k=1}^{K^*+1}$ define the **contextual** aspect for the anomaly detection step implemented in Section 5.2.3.

5.2. Methods

We present the **three steps** methodology that leads to detecting contextual anomalies.

5.2.1. Temporal change point detection

The method for estimating K^*, $(\tau_k^*)_{k=1}^{K^*+1}$ and $(\boldsymbol{\theta}_k^*)_{k=1}^{K^*+1}$ for the coarse grained resolution $\overline{\mathcal{D}}$ is fully described in Section 4.3 of Chapter 4. We denote $\widehat{K}$, $(\widehat{\tau}_k)_{k=1}^{\widehat{K}+1}$ and $\widehat{\theta}$ the resulting estimated parameters. The estimation of these parameters is the **first step** of the whole monitoring procedure.
In the data used in this work, it should be noted that 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).

5.2.2. Spatial clustering

We propose to use the graph $G = (V, E)$ to build spatial aggregates and to assess homogeneity at this aggregated level. This corresponds to clustering the stations using the graph structure and thus environmental information.
The goal is to successfully create a global partition in M cluster of stations in the presence of R non connected components in the graph. This raises the question of how to dispatch these M clusters among the non connected components. We present two different methods to solve this problem.

Criterion choice

In order to solve the problem of (spatial) segmentation of R components $(\mathcal{K}_1, \ldots, \mathcal{K}_R)$ into M clusters, we introduce the following notations for a clustering $\mathcal{P}$:

- C_m^r the m-th cluster located in $\mathcal{K}_r$.
- M_r the number of clusters in $\mathcal{K}_r$.
- $\mathcal{P}$ the clustering is defined as $(C_1^1, \ldots, C_{M_1}^1, C_1^2, \ldots, C_{M_2}^2, \ldots, C_1^R, \ldots, C_{M_R}^R)$
- $Q(\mathcal{K}_r, C_m^r) = \dfrac{1}{|C_m^r|} \displaystyle\sum_{v_i, v_j \in C_m^r} d_{ij}^2$ the inertia of cluster C_m^r, where d_{ij}^2 is the square of the shortest path distance in the graph G between vertices v_i and v_j, and $|C_m^r|$ denotes the cardinality of set C_m^r.
- $U_r(M_r) = \displaystyle\min_{(C_m^r)_{m=1}^{M_r}} \sum_{m=1}^{M_r} Q(\mathcal{K}_r, C_m^r)$ the best partition (in the sense of minimal inertia) of $\mathcal{K}_r$ into M_r clusters.
- $S(l, m) = \displaystyle\min_{(M_r)_{r=1}^{l} \text{ such that } \sum_{r=1}^{l} M_r = m} \sum_{r=1}^{l} U_r(M_r)$ which is the best partition of the l first components into a total number of m clusters.

With these notations, we can write the global inertia of clustering $\mathcal{P}$ as:

$$\mathcal{C}(\mathcal{P}) = \sum_{r=1}^{R} \sum_{m=1}^{M_r} Q(\mathcal{K}_r, C_m^r) \tag{5.4}$$

For each $\mathcal{K}_r$, the stations are clustered using Ward hierarchical clustering method. Since Ward method is deterministic, $U_r(M_r)$ the best partition of $\mathcal{K}_r$ in M_r clusters is entirely determined by M_r.
Therefore, It is sufficient to find the optimal number of segments M_r in each component to minimise the criterion (5.4). More precisely, given a total number of clusters M, we are looking for:

$$(\widehat{M}_r)_{r=1}^{R} = \arg \min_{(M_r)_{r=1}^{R} \in \mathbb{N}^R | \sum_{r=1}^{R} M_r = M} \sum_{r=1}^{R} U_r(M_r) \tag{5.5}$$

We present two methods to search for these values in the next section.

Search methods for the clustering

In the following, we give two algorithm descriptions and pseudo-codes to find the optimal number of segments M_r in each connected component of the graph $G = (V, E)$. Note that, in practice, $U_r(M_r)$ is computed with Ward hierarchical clustering technique implemented the `R` package `hclust`.

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 reduces the most the inertia of the clustering defined in Equation (5.4). This method is presented in Algorithm 5.

Algorithm 5 Clustering with greedy method:

input : the station graph $G = (V, E)$, the known partition into non connex components $(\mathcal{K}_1, \ldots, \mathcal{K}_R)$, a total number of clusters M

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

for $m = 1$ to $M - R$ **do**
  $score \leftarrow (0, \ldots, 0)$ vector of size R
  for $r = 1$ to R **do**
    $M_{opt}(r) \leftarrow M_{opt}(r) + 1$
    $score(r) \leftarrow \sum_{r=1}^{R} U_r(M_{opt}(r))$
    $M_{opt}(r) \leftarrow M_{opt}(r) - 1$
  end for
  $pos \leftarrow \arg\min_{r \in \{1,\ldots,R\}}(score)$
  $M_{opt}(pos) \leftarrow M_{opt}(pos) + 1$
end for
for $r = 1$ to R **do**
  built the optimal partition of $\mathcal{K}_r$ with $M_{opt}(r)$ clusters using `hclust`.
end for

2. **A clustering method based on dynamic programming:** this approach is derived from Hébrail et al. (2010). This paper shows that is possible to cluster the stations graph into M segments in presence of R non connected components. Since (5.4) is an additive criterion, it is possible to distribute the M segments among the R components in an optimal way using dynamic programming. Our context is a little bit simpler than Hébrail et al. (2010) since the R components are already known and doesn't have to be estimated. This method is presented in Algorithm 6.

Selecting the optimal number of clusters

To select the final clustering in the hierarchy, we use the same elbow method heuristic as in Algorithm 2. For several values of $M \in \{R + 1, \ldots, M_{max}\}$, the inertia of the clustering is plotted against the corresponding number of clusters M. We look for the number of clusters M^* that minimizes the sum of squares of a two part linear models respectively fitted on the $M \geq M^*$ and the $M \leq M^*$.

Algorithm 6 Clustering by dynamic programming:

input : the station graph $G = (V, E)$, the known partition into non connex components $(\mathcal{K}_1, \ldots, \mathcal{K}_R)$, a total number of clusters M

for $r = 1$ to R **do** :
  Use `hclust` to compute $U_r(m)$ for all $m \in \{1, \ldots, M - R + 1\}$
end for
for $m = 1$ to $M - R + 1$ **do** :
  $S(1, m) \leftarrow U_1(m)$
end for
for $l = 2$ to R **do** :
  for $m = l$ to M **do** :
    $W(l, m) \leftarrow 1$
    $S(l, m) \leftarrow S(l - 1, m - 1) + U_l(1)$
    for $u = 1$ to $m - l + 1$ **do**
      if $S(l - 1, m - u) + U_l(u) < S(l, m)$ **then**
        $W(l, m) \leftarrow u$
        $S(l, m) \leftarrow S(l - 1, m - u) + U_l(u)$
      end if
    end for
  end for
end for
$M_{opt} \leftarrow (\mathrm{NA}, \ldots, \mathrm{NA})$, vector of size R
$M_{opt}(R) \leftarrow W(R, M)$
$left \leftarrow M - W(R, M)$
for $r = R - 1$ to 1 **do**
  $M_{opt}(r) \leftarrow W(r, left)$
  $left \leftarrow left - W(r, left)$
end for
for $r = 1$ to R **do**
  built the optimal partition of $\mathcal{K}_r$ with $M_{opt}(r)$ clusters using `hclust`
end for

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. Newman (2003)) that justify the use of techniques such as maximal modularity clustering see e.g. Fortunato (2010). On the contrary, simpler approaches that interpret shortest paths weights as dissimilarities should be sufficient see e.g. Schaeffer (2007).
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, unlike what was previously done in Laroche et al. (2022a), the spatial clustering is independent of the temporal context $[\widehat{\tau}_k, \widehat{\tau}_{k+1}]$. The clustering is performed on the graph $G = (V, E)$ composed of all stations available in the data. This spatial clustering constitutes the **second step** of the whole monitoring procedure.

5.2.3. Anomaly detection

Conditionally to a given temporal segment $[\widehat{\tau}_k, \widehat{\tau}_{k+1}]$ in the time segmentation selected in the first step (see Section 5.2.1), we can now compare the spatial clusters detected in Section 5.2.2. From this comparison, we derive an anomaly detection procedure. This is the **third step** of the procedure. We identify two cluster characteristics that are likely to help detecting anomalies. The first criterion is based on the within cluster concentration distribution. The second criterion assesses the homogeneity between the station concentration profiles in a cluster.

Criterion based on within cluster concentration distribution

For each spatial cluster C_m^r, we estimate the parameters of the distribution Q (see Section 5.1.3) on all measurements collected in C_m^r during the selected temporal segment.
From these parameters, we compute a statistic, denoted $\bar{I}_m$, used as a proxy for the intensity of the measurements (see Section 5.4.3 for an example).
In the example where Q is the Weibull distribution, the scale parameter λ gives an indication on the mean since it can be expressed as:

$$\lambda \Gamma\left(1 + \frac{1}{\sigma}\right) \tag{5.6}$$

In the case where all clusters share the same shape parameter σ, λ is a good indicator of the intensity of the measurements in a selected cluster. We would then set $\bar{I}_m = \widehat{\lambda}_m$.
In other words, for this criterion, anomalous clusters are detected by simply pooling the measurements of all stations in each cluster to estimate the local use of the substance and detect large rates.

Criterion based on station heterogeneity

This criterion aims at quantifying the heterogeneity of the measurements provided by the stations in a given spatial cluster during the time period defined by the selected temporal segment.

As pointed out previously (see Chapter 2), 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 sampling time in common.

For this reason, we propose to use the Wasserstein w_1 distance (Villani, 2009) adapted for left censored variables. For two empirical distributions on $\mathbb{R}$, it is expressed as the L^1-distance between their cumulative distribution functions and is therefore simple to compute.

The Wasserstein distance was preferred to the Kolmogorov-Smirnov or the Jensen-Shannon metric. It has the advantage of including both the differences between the probabilities of observing different values and the metric structure of the space of possible concentration values ($\mathbb{R}^+$). This is illustrated by a simple simulated example in Figure 5.1.

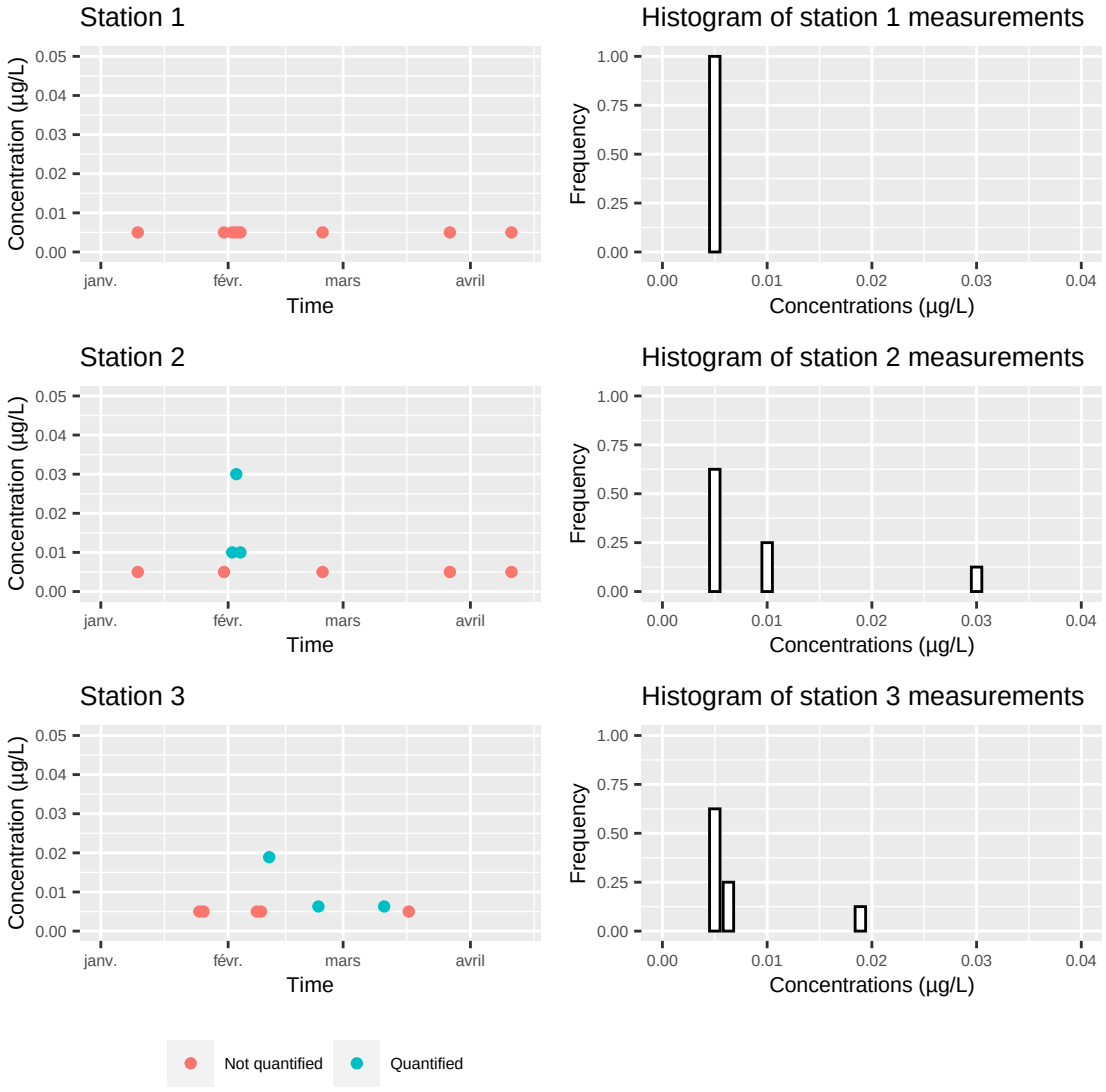

Figure 5.1: Example of three stations data. The data were simulated.

Three monitoring stations with different concentration behaviours are represented in Figure 5.1. These different behaviours are evident in both the temporal plot and the histograms.
The Kolmogorov-Smirnov distance between stations 1 and 3 is equal to the Kolmogorov-Smirnov distance between stations 1 and 2. The same holds for the Jensen-Shannon distance. These distances 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.
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:

$$w_1(F_n, G_n) = \int_{\mathbb{R}} |F_n(x) - G_m(x)| \mathrm{d}x \tag{5.7}$$

To adapt to the case of left censored observations, we make the assumption that the real values under the censoring threshold are uniformly distributed and we modify the c.d.f accordingly to this assumption (see Figure 5.2). This assumption seems reasonable since the samples size for a single station is usually very small.
Figure 5.2 illustrates the changes it implies on the empirical c.d.f.. Note that we did not represent the case where a single station have several censoring threshold values (because it changed its equipment for example). This case did not occur when handling real data.

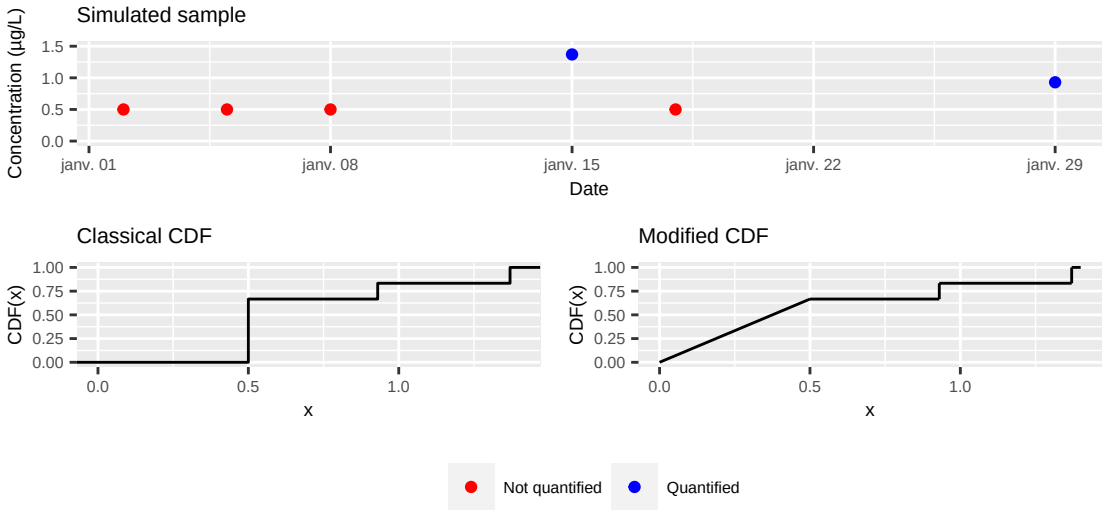

Figure 5.2: Example of modified c.d.f. for the Wasserstein distance.

Denoting $w_1(\mathbf{y}_i, \mathbf{y}_j)$ the adapted empirical 1-Wasserstein distance between the data of stations v_i and v_j, this criterion can be written as

$$\bar{W}_m = \frac{1}{|C_m|(|C_m|-1)} \sum_{1 \leq j \leq |C_m|} \sum_{1 \leq i \leq |C_m|, i \neq j} w_1(\mathbf{y}_i, \mathbf{y}_j). \tag{5.8}$$

Intuitively, when stations with different concentration profiles are present in C_m, $\bar{W}_m$ takes high values. As pointed out in the chapter introduction, we expect a low (or moderate) heterogeneity level within each cluster. Therefore, anomalous clusters are associated with high values of $\bar{W}_m$.

Multi-criteria analysis

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 Kießling (2002), we say that $X_m = (\bar{W}_m, \bar{I}_m)$ 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. Pareto front and levels are evaluated using the Skyline algorithm Borzsony et al. (2001); Endres et al. (2015).
Note that we do not decide if a cluster is anomalous or not per say, we rank them according to two criteria summarising anomalous behaviours. The actual abnormality of these clusters must be assessed by expert knowledge.

5.3. Data presentation

The methodology introduced in the above sections will be illustrated next using a case study on the prosulfocarb concentration National Center for Biotechnology Information (n.d.) 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 ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) (n.d.); Agriculture & Environment Research Unit (AERU) at the University of Hertfordshire (n.d.).

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) 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 5.3).

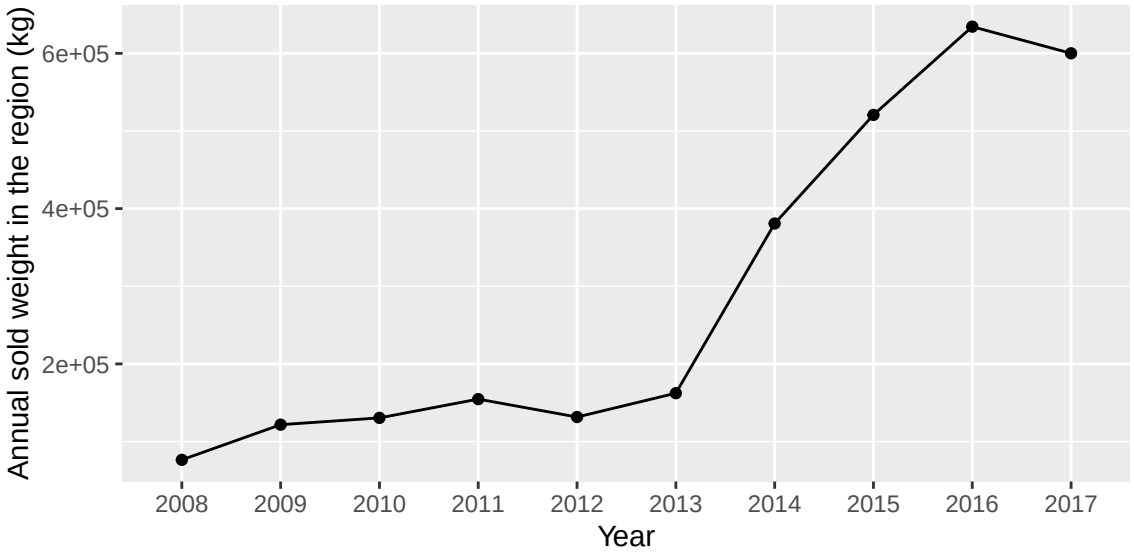

Figure 5.3: Prosulfocarb sales between 2008 and 2017 in the Centre-Val de Loire region.

This region is also characterised by high concentrations of prosulfocarb target crops such as the Beauce plains (see Figure 5.13). 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 made available by the French Biodiversity Agency Office français de la biodiversité (n.d.) and we provide the link of the query[1] with which we selected the data.
This query led to a data set $\mathcal{D}$ comprising $I = 420$ 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.
Among the 14,203 recorded measurements during the period of interest, only 14.11% were above the quantification limit. Figure 5.4 shows the distribution of the number of measurements per 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.
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 5.5. The aggregated series contains $n = 2{,}150$ values, among which 22.51% are quantified.

[1]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`

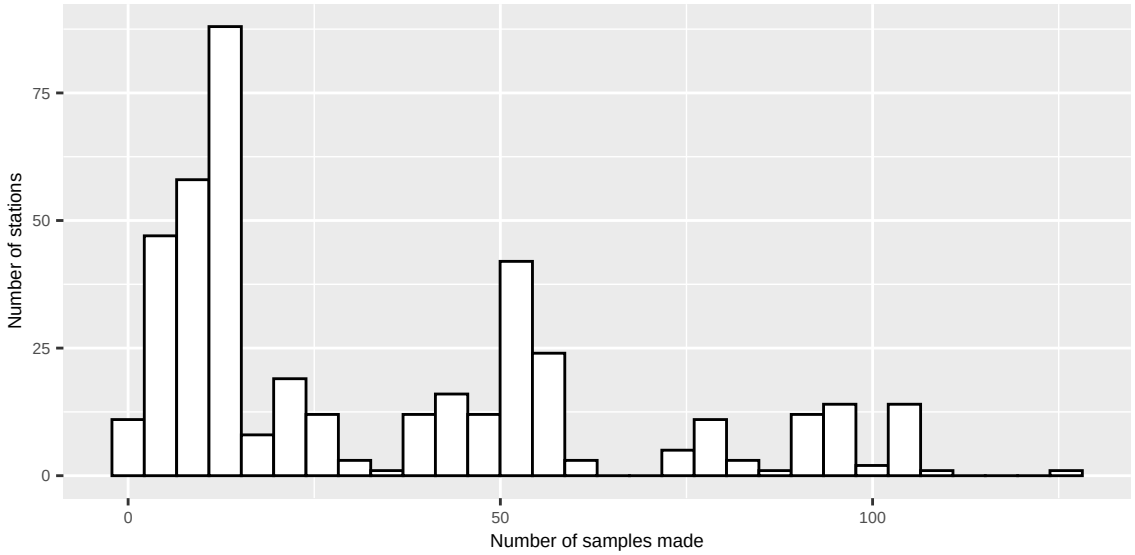

Figure 5.4: Distribution of the number of measurements per station.

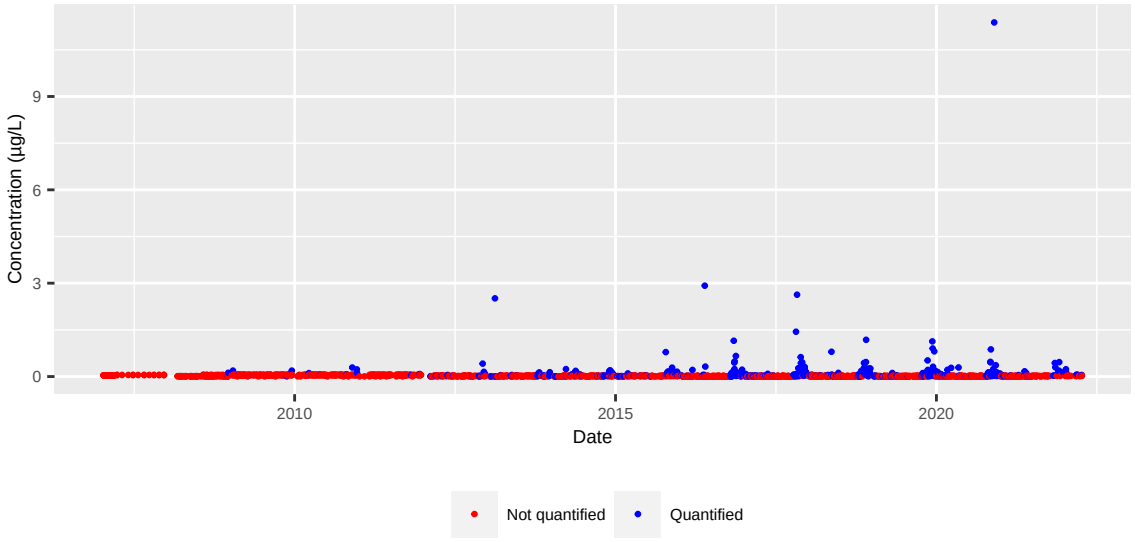

Figure 5.5: Plot of daily maximum concentrations.

One may note here that despite the aggregation process, the coarse series remains irregularly sampled, and that no measurementt were made for apprrox. two thirds of the days included.

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) Institut National de l'Information Géographique et Forestière (n.d.) 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 $\mathcal{R} = (S, H)$, where the vertices S are made of the connecting points between the river sections, and the edges H contain all sections. 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 $\mathcal{R}$, 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 5.6 displays the graph of all stations that made at least one measurement during the observation period. The graph is not fully connected and exhibits 9 distinct connected components.

5.4. Results

The results of the **three steps** leading to the prosulfocarb concentration monitoring are presented:

1. We first present the change-point detection results on the daily maximum concentration levels of prosulfocarb.

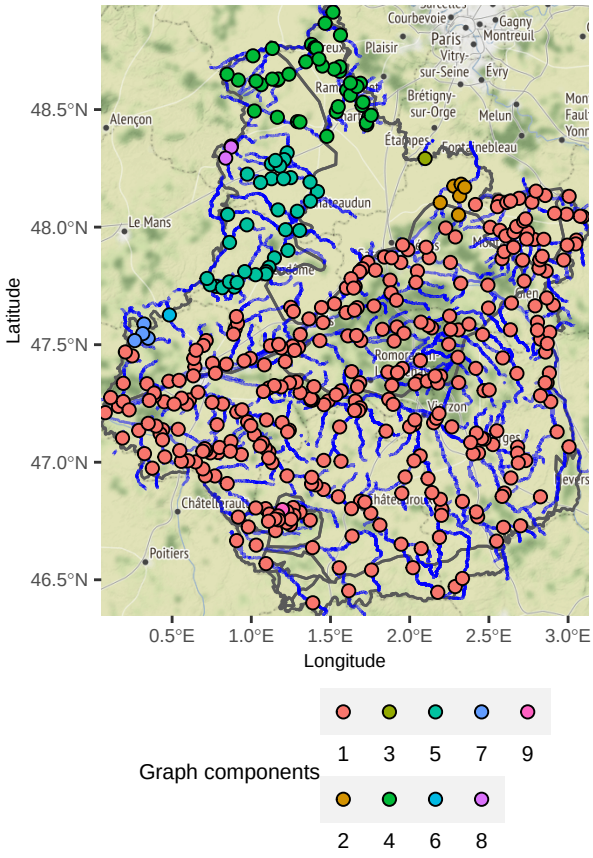

Figure 5.6: Map of the non connex components in the station graph.

2. Then we show the results of clustering on the station graph using the river system of Centre-Val de Loire region.

3. We compute the Pareto front of these clusters in a selected time period derived from the first step.

Even though we chose to present it in this order, note that the first and second steps are independent.

5.4.1. Temporal segmentation

First, the coarse-grained time series $\overline{\mathcal{D}}$ in Figure 5.5 is segmented using the change-point detection procedure described in Section 4.3.
We fit a left censored Weibull distribution in the change-point model with parameters $\theta^* = (\lambda^*, \sigma^*)$. This was motivated by the observation of the data.
We suppose that the constant parameter over the segments is σ^*. The scale parameters $(\lambda_k^*)_{k=0}^{K^*}$ are supposed to change values at each of the K^* change-points.
In this model, the log-likelihood of a segment $y_{u:v}$ can be written:

$$\mathcal{L}(y_{u:v}, \lambda, \sigma) = \sum_{i=u}^{v} \log\left(1-\exp(-(\lambda y_i)^\sigma)\right)\mathbb{1}_{y_i=a_i} + \sum_{i=u}^{v}\left(\log(\lambda\sigma)+(\sigma-1)\log(\lambda y_i)-(\lambda y_i)^\sigma\right)\mathbb{1}_{y_i>a_i}, \tag{5.9}$$

Assuming σ^* is known, we can derive the cost function from (5.9) and the parameter estimators $\widehat{K}$, $(\widehat{\tau}_k)_{k=0}^{\widehat{K}}$ and $(\widehat{\lambda}_k)_{k=0}^{\widehat{K}}$ following the approach described in Section 4.1.
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.
If we now suppose σ^* to be unknown, we use the estimation procedure of section 4.3. The hyper parameters used in the estimation procedure are the following:

- The value of θ_{max} (discussed in Section 4.2) is set to 10^6 which is superior to the value we would obtain using (4.11) thus ensuring the identifiability of the scale parameters.

- The minimum segment size used in the PELT search method is set to 50. The censoring rate of $\overline{\mathcal{D}}$ is 77.48%. We know that this minimum segment size should provide satisfactory results from the experiments of Section 4.4.

- The penalty range explored is set to $\left[\beta_{min} = \frac{\log(n)}{5}, \beta_{max} = 5\log(n)\right]$. Initial runs of the estimation procedure with these penalty values result in two segmentations with respectively 1 and 30 change-points. Exploring segmentation results that have a number of change points within that range seems reasonable.

- Since we provide an illustration of a segmentation in this section, we had to select a segmentation result to present. We used the elbow heuristic to select this result knowing that it does not necessarily provide an optimal segmentation.

The results of the procedure are presented in Figure 5.7. 11 different segmentations were uncovered using CROPS. The associated number of change points of these segmentations range from 1 to 30. No segmentation with K change points with $K \in \{2, \ldots, 12\}$ was uncovered. This could highlight the presence of several obvious change-points that cannot be missed. The associated estimated values of σ^* range between 0.31 and 0.39. This confirms that the data has a heavier tail than an exponential distribution (σ=1), and that the the use of Weibull distributions for our data is appropriate.

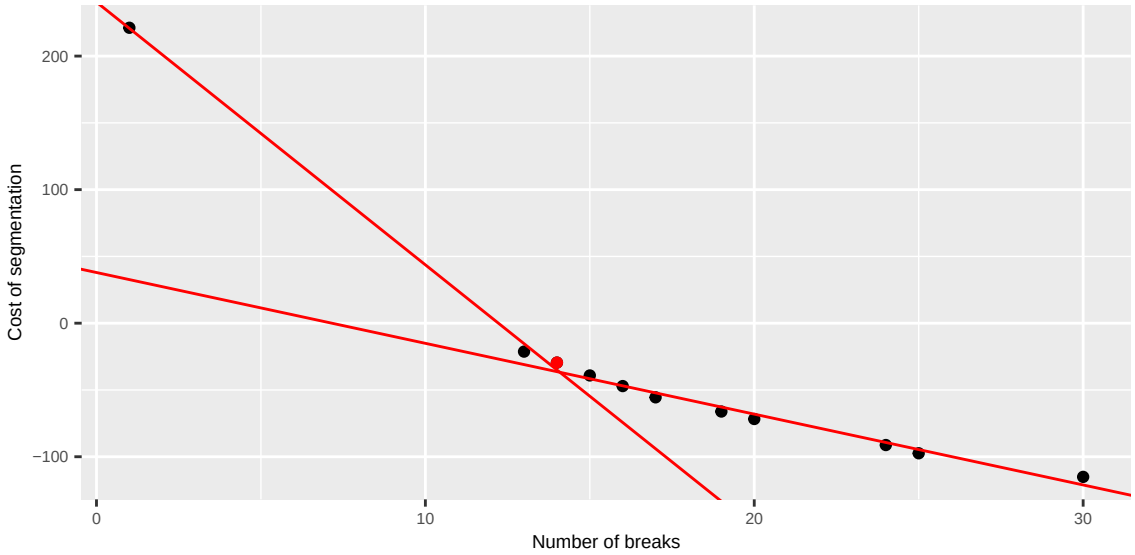

Figure 5.7: Elbow method selecting the optimal segmentation of the full signal $\overline{\mathcal{D}}$.

According to Figure 5.8, 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 emerge. They correspond respectively to the periods of intensive usage of prosulfocarb and to 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 5.3.

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 of the temporal segmentation. Algorithm 6 based on dynamic programming was used to create the partition of the station graph.

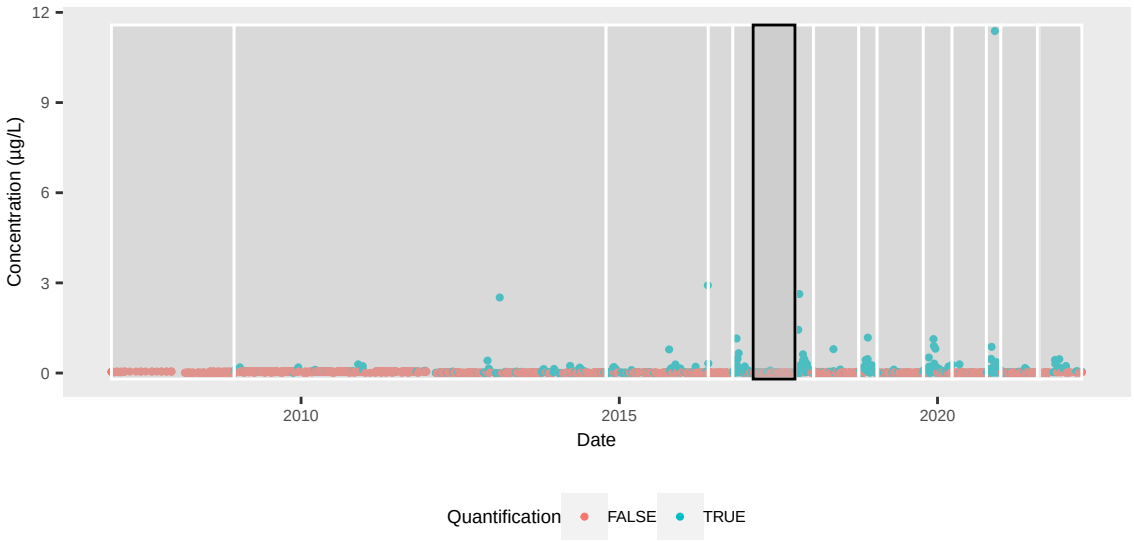

Figure 5.8: Segmentation found selected by the elbow heuristic in Figure 5.7.
The dates of the breaks are : December 15, 2008; October 20, 2014; May 26, 2016; October 14, 2016; February 8, 2017; October 9, 2017; January 22, 2018; October 8, 2018; January 23, 2019; October 14, 2019; March 25, 2020; October 8, 2020; December 30, 2020; August 8, 2021.The black rectangle corresponds to the selected temporal segment in Section 5.4.2.
The estimated values of σ^* and λ^* are $\widehat{\sigma} = 0.36$ and
$(\widehat{\lambda}_k)_{k=0}^{14} = (10^6, 521, 163, 3180, 23, 798, 13, 700, 20, 1139, 35, 1470, 12, 794, 77)$

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 6 and 35. The minimum number of clusters explored is equal to the number of connected components composed of more than one station in the graph.
The optimal number of clusters was selected using the elbow method applied to the inertia curve. According to this heuristic, illustrated in Figure 5.9, the best solution corresponds to a 15-clusters configuration. The spatial segmentation is illustrated in Figure 5.10.

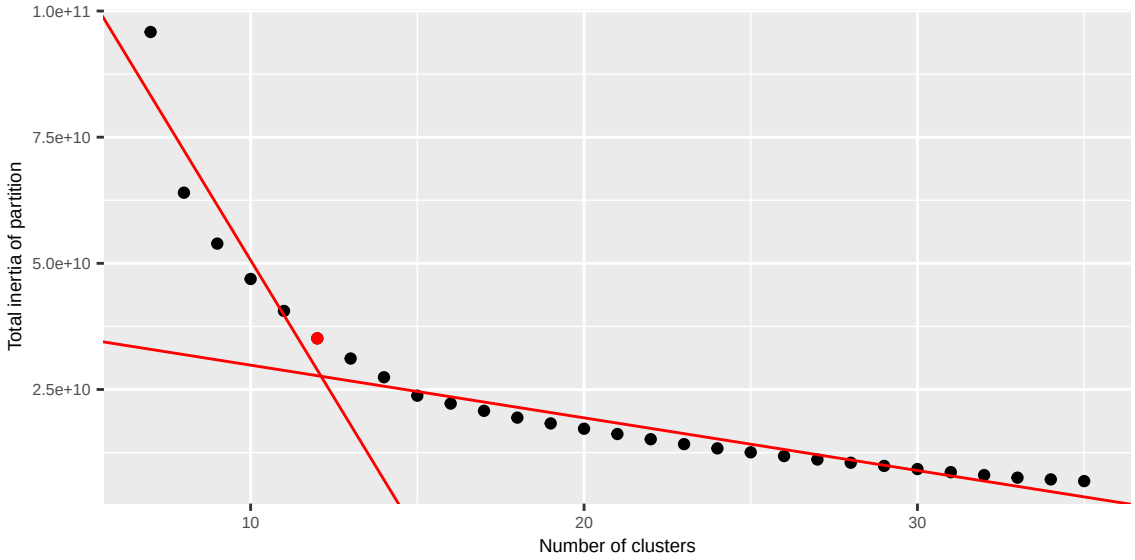

Figure 5.9: Elbow method for the spatial clustering.

As expected the biggest component in Figure 5.6 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.
Once a clustering result is selected, we want to check the relevance of the homogeneity assumption formulated in Section 5.2.2. Conditionally to a temporal segment, the homogeneity is assessed comparing the global average empirical pairwise Wasserstein distance of all active stations and the within cluster average empirical pairwise Wasserstein distance.
We test in practice the homogeneity assumption on the selected clustering in Figure 5.9.
An off-peak period, spanning between February 8, 2017 and October 9, 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 5.8.
We proceeded in two steps:

1. We pooled all samples made during this specific period of time. From all these samples, we can identify the active clusters during this period, they were 13 out of 15.

2. For all active clusters, we computed the within average empirical pairwise Wasserstein distance.

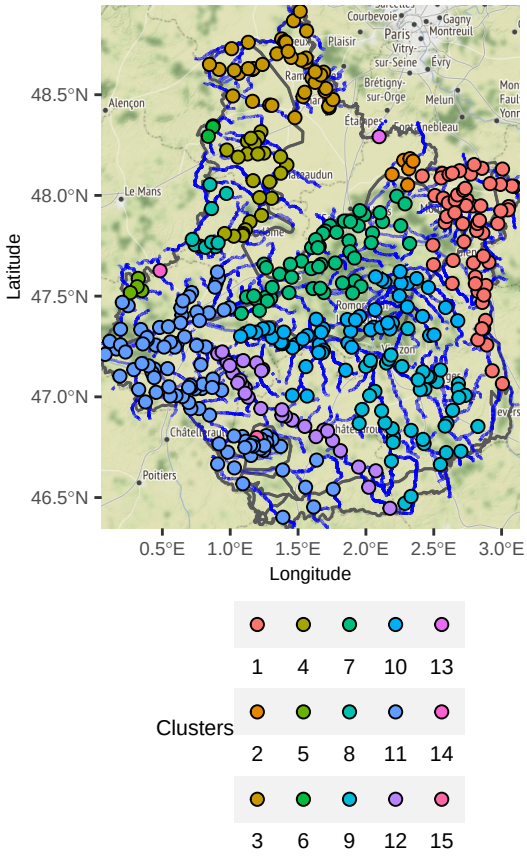

Figure 5.10: Map of geographical clusters.

We observe that for 10 clusters out of 13, this indicator is less than 0.0015, whereas the global average pairwise Wasserstein distance for the 420 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 Wasson et al. (2002). The borders of those regions are drawn in grey in Figure 5.10. This ensures that the substances will have homogeneous dispersion properties on these clusters (see cluster 7).

5.4.3. Anomalous cluster identification

We now focus on locating spatial patterns during time segments identified in Section 5.4.1. We select a segment corresponding to an off-peak period since it should be abnormal to find some concentration levels in these periods. In the rest of this case study, we selected the segment highlighted in black in Figure 5.8 (from February 8, 2017 to October 9, 2017). This introduces a context to the anomaly detection: a global non use of the substance.

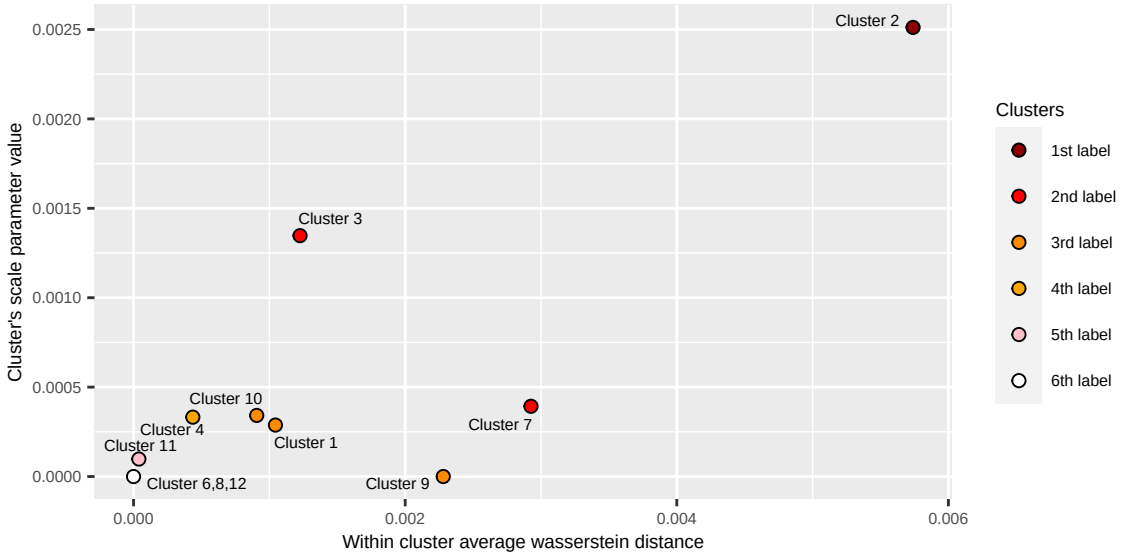

Figure 5.11: Clusters pareto front.

Following the methodology proposed in Section 5.2.3, the scaling parameter λ_k of the aggregated data of each spatial cluster found in Section 5.4.2 was estimated. Since we used an alternative parametrization in (5.9) for calculation purposes, the mean of the Weibull in our parametrization writes as:

$$\frac{1}{\lambda}\Gamma\left(1+\frac{1}{\sigma}\right) \tag{5.10}$$

Therefore, we set statistics $\bar{I}_m$ to $1/\hat{\lambda}_m$

The Pareto front involving the two descriptors $\bar{W}_m$ and $\bar{I}_m$ was computed. It led to the cluster ranking displayed in Figure 5.11 using the *rPref* package Roocks (2016).
We recall that the selected time segment corresponds to a period of non-use of prosulfocarb. Thus, 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 5.11:

- **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.

The Pareto front level is not uniformly distributed in the region. Figure 5.12 displays the Pareto front levels on the station map. The anomalous clusters listed above are located in the north and east of the region. Their location could be related to the agricultural practices and land use. It is interesting to note that there is a high concentration of barley crops in Figure 5.13 near the location of the most anomalous cluster.

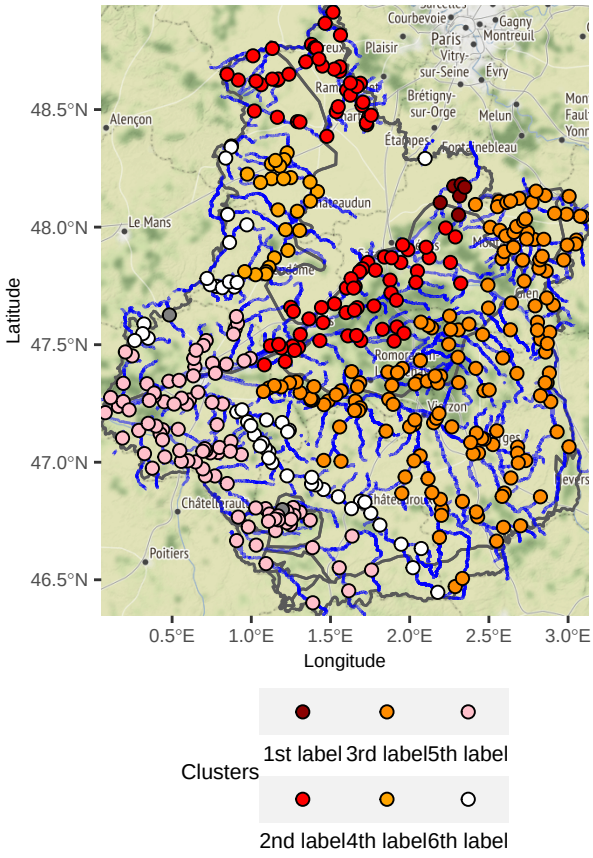

Figure 5.12: Mapped pareto front.

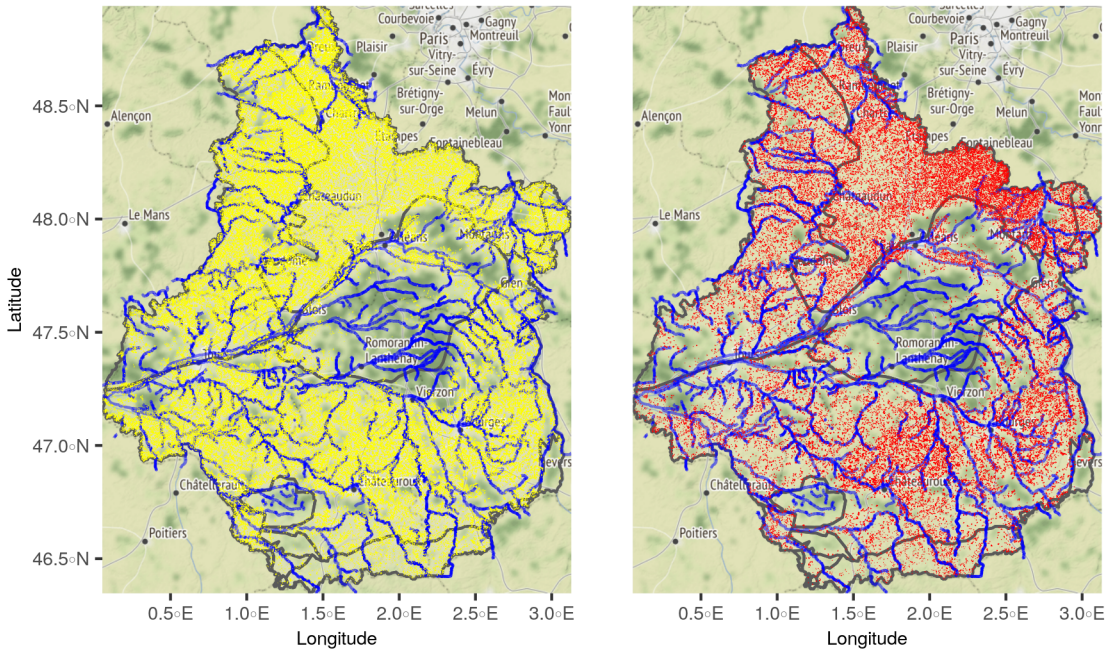

Figure 5.13: Wheat (in yellow) and barley (in red) crops location in Centre-Val de Loire.

5.5. Chapter summary

This Chapter presents the statistical analysis carried out to monitor the concentration data. The procedure is divided into three steps. The first two steps are independent and deal with the characteristics of the concentration data presented in Chapter 2. The first step is the temporal change-point detection method developed in Chapter 4, adapted for censored data. Homogeneous time segments are found in the aggregated time series of daily maximum concentrations observed in the study region. This step leads to the identification of the use regimes of a substance in the region. The second step deals with spatial heterogeneity. It is assumed that the environment in which the substance diffuses explains most of the spatial heterogeneity of the concentration values. The second step is to cluster the stations according to the structure of the environment. The structure of the environment represents a proxy that should capture geographical areas where the concentration values should be homogeneous. This step of clustering is done with Ward's hierarchical clustering. The final step is to identify anomalous clusters in the context of a temporal segment. This can be done using multi-criteria optimisation. Each cluster found in the second step is characterised by two criteria calculated from all concentration measurements of the stations forming the cluster. The first criterion evaluates the presence of anomalous stations in the cluster, the second criterion evaluates the overall level of concentration in the cluster. The Pareto front calculated in the multi-criteria optimisation makes it possible to characterise the degree of anomaly of a cluster
The whole procedure is illustrated by the prosulfocarb concentration in the Centre-Val de Loire region. The main result is that the levels of the Pareto front are not evenly distributed over the area of the region. This makes it possible to identify 'epicentres' of substance use in the region. In this chapter, the results are only presented for one temporal segment. In Chapter 6 an interactive tool is presented that allows a user to examine the results in all segments of a segmentation as well as multiple segmentation results. This is a monitoring tool that combines statistical inference with the visualisation techniques mentioned in Chapter 2.

6. Software development

Contents

In this Chapter, we present the design and the implementation of the procedure developed to detect spatio-temporal anomalies described in Chapters 4 and 5.
The implementation was done in two steps. First, for a selected pollutant during a selected period in a selected geographical area, an R script implements the preprocessing of raw data as well as the precomputation of change-point detection in concentration data as described in Chapter 4 and the spatial clustering of measuring stations as described in Chapter 5. This script can handle the concentration in surface waters of any substance in any region in France. This preprocessing step is described in Section 6.1, together with the format of the produced output file.
The file produced in the first step is then fed to an Rshiny application, described in Section 6.2. This application is intended for the ANSES experts in order to analyse pollutant concentration data in surface waters, therefore the interface and user manual are written in French.
So far, the application can handle Prosulfocarb concentration data only, however it can very easily be extended to other substances in future developments. Three options are available for the user, implemented as three tabs in the main menu. The Home tab, presented in section 6.2.1, provides an overview of the available information on the substance under consideration (here the Prosulfocarb), as well as basic visualization of the input data. The Detection tab allows the user to perform the anomaly detection procedure described in Chapter 5. The corresponding implementation is described in Section 6.2.2. The third tab allows accessing the user manual of the application. This document, intended for users that do not necessarily have a strong mathematical background, is available in Appendix B.2.
The general organisation of the preprocessing, precomputation and application is represented in Figure 6.1.

6.1. Data preprocessing and precomputation

The raw data to be processed are the three following files that the user has to extract from public databases:

- A table of the concentration measurements in a given geographical area and during a given time period (in the case of surface waters substances, such tables can be extracted from NAIADE[1] database in .csv file format).

- A table of the station locations (.csv file format extracted from NAIADE database) in Lambert 93 Coordinates Reference System (CRS).

- For each administrative region ('département' in French) in the geographical region of interest, a file providing the coordinates and direction of the hydrographic segments (shapefile format, extracted from the BDTOPO[2] database)

An `R` script was developed to preprocess these data and precompute time-segmentation and spatial clustering. The data preprocessing steps consist of:

[1] see *Web resource: Naiades portal* (n.d.)
[2] see *Web resource: BDTOPO database* (n.d.)

- The construction of the station graph, according to the method described in Section 5.3.2. Its vertices are the stations and its edges indicate a connexion between two stations through the hydrographic network. The edge weights are the length of the shortest existing path of water linking two stations. The graph is stored as a `igraph` object (Csardi & Nepusz, 2006).

- The construction of a table for measured concentrations, storing the sampling date, the measurement, a boolean indicating if this value corresponds to a quantified value, the station ID and the LOQ value. All values are converted to μg/L. The table containing the daily maximum values time series presented in Section 5.1.2 is constructed from this table.

- The construction of a regional shapefile merging all "département" shapefiles, converted to Lambert 93 CRS.

For the precomputation step, the change-point detection method proposed in Chapter 4 is implemented for a Weibull distribution. The cost function evaluation and the PELT algorithm are coded with `Rcpp` for a faster execution, while CROPS algorithm is coded in `R`.
The following quantities are calculated:

- The change-point detection in the daily maximal concentration signal in the geographical area of interest. The CROPS algorithm is trained for a penalty range $[\beta_{min}, \beta_{max}]$ defined in Section 5.4.1. The significant penalties selected by CROPS and the associated segmentation results are stored in a list.

- The connected components of the station graph, identified using the available `components` function in the `igraph` package.

- The spatial clustering of the station graph is performed using the clustering methods presented in Algorithm 6. We did not use the elbow heuristic to select a unique number of clusters because we wanted to offer the experts the opportunity to explore different values. This is why several clustering results are stored, with their respective number of clusters ranging between a minimal and maximal value. The heuristic used to determine the minimal and maximal number of clusters is presented in Appendix B.1.

The output file of this preprocessing and precomputation procedure is a `.Rdata` file containing:

- The detailed table containing all concentration measures for all stations.

- The table with all daily maximum measurement information built in the preprocessing step.

- A list with all temporal segmentation results and details obtained by CROPS (parameter values, cost, penalty values etc...).

- A table with all station coordinates in WGS84 CRS and with all clustering membership results and connected component labels.

- A table with all hydrographic segment coordinates in WGS84.

- A table with the geographical area border coordinates in WGS84.

- A table with the border coordinates of the hydro ecoregions in WGS84.

6.2. Rshiny application

As mentioned in the introduction, this application aims at visualizing the preprocessed data (subsection 6.2.1) as well as the results of the anomaly detection procedure described in Chapter 5 (Section 6.2.2). When the application is started, the `.Rdata` file generated by the preprocessing step in Section 6.1 from the raw concentration and geographic datasets is automatically uploaded. According to Figure 6.1, this `.Rdata` file will be referred as the input file in the rest of this section.

6.2.1. Basic visualisation

In this part, which is accessible through the "Home" tab, the user can visualize across three panels information contained in the input file.
In the first panel, the following basic information is provided:

- The name of the substance corresponding to the input file uploaded.
- The dates corresponding the time span in the input data.
- The geographical area corresponding to the input file uploaded.
- The total number of measurements collected over the times period.
- The number of active stations over the time period.
- The percentage of quantified concentration results among these measurements.
- The number of days where at least one sample was recorded. This number is also the number of daily maximum concentrations.
- The percentage of quantified daily maximum concentrations.

The second panel (Figure 6.2) displays the plot of the daily maximum concentrations in the geographical area of interest, which allows the experts to visualize the temporal trends and seasonality at a glance, before having any segmentation information that could influence their interpretation.
The third panel (Figure 6.3) displays the map of the active stations during the period, together with the river network in the area. Based on the preprocessing in section 6.1, the stations are coloured according to the hydrographic graph connected component they belong to.

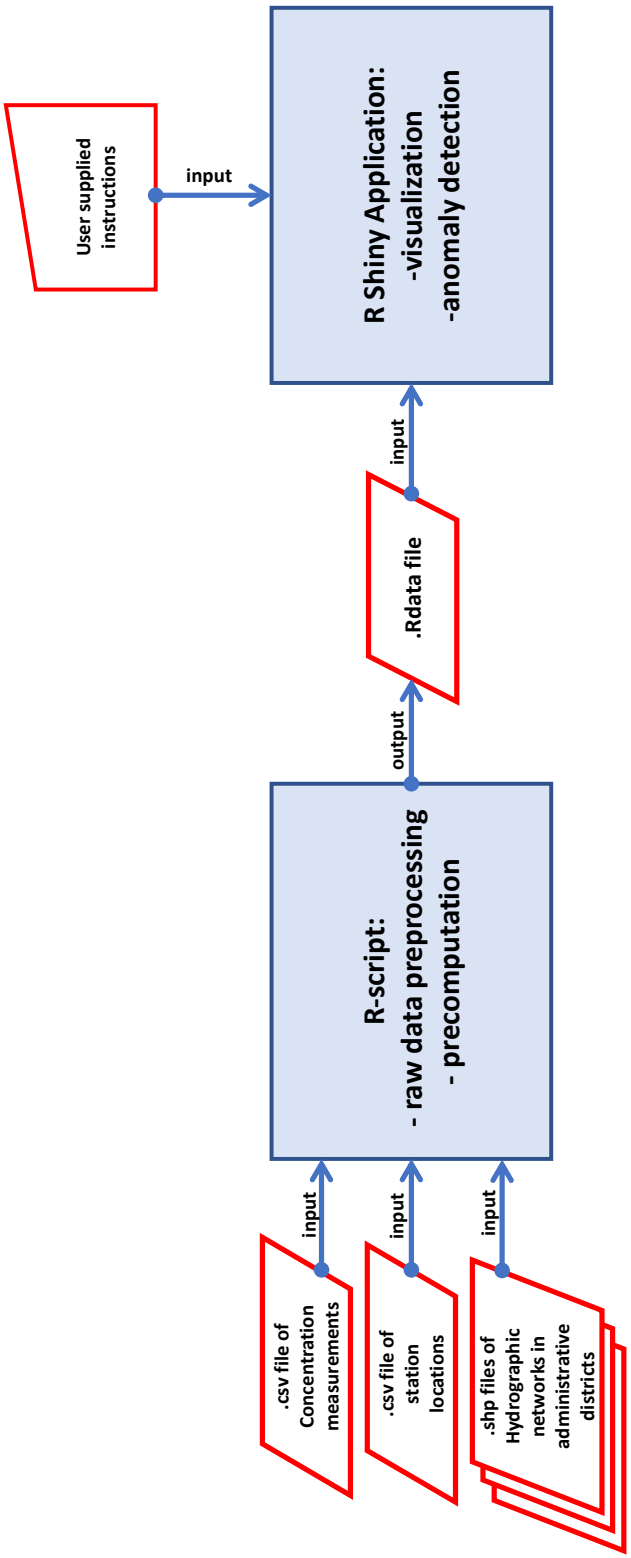

Figure 6.1: Overall organisation describing the articulation between the preprocessing, the precomputation and the application, the inputs needed and the outputs produced.

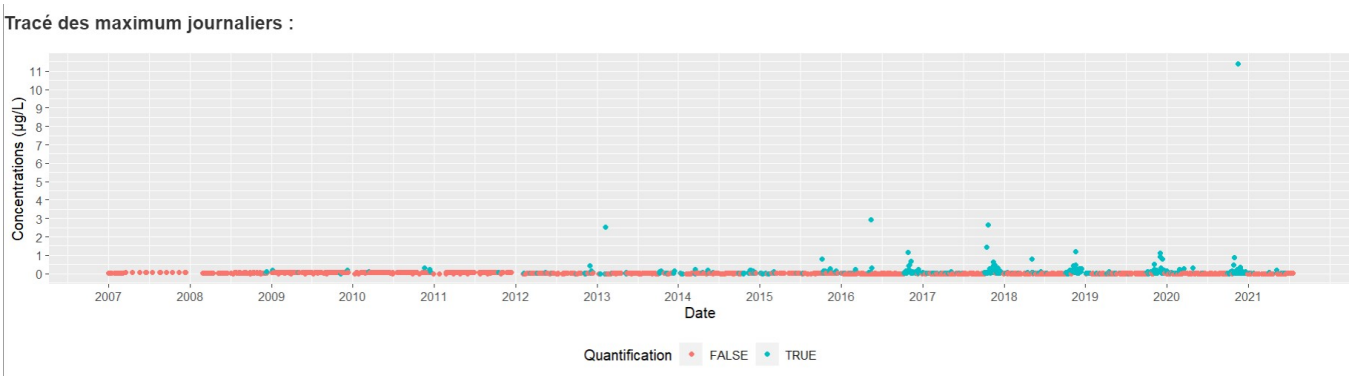

Figure 6.2: Global temporal presentation.

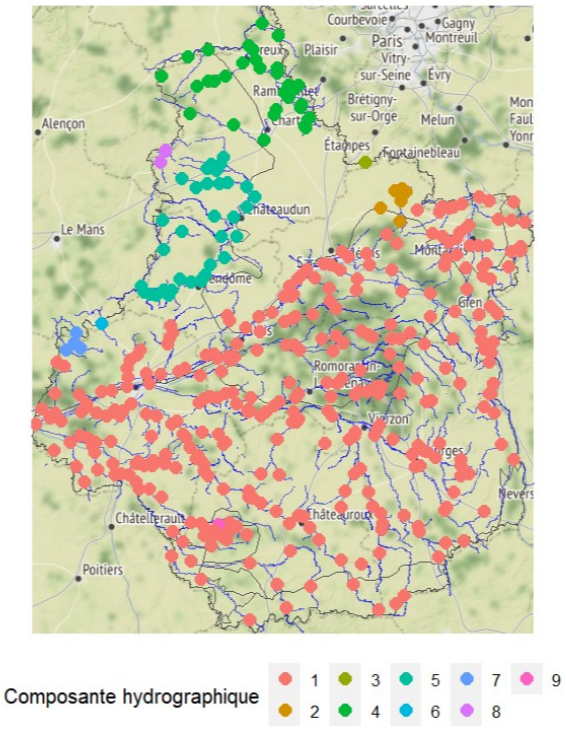

Figure 6.3: Global geographical presentation.

6.2.2. Anomaly detection

This part is accessible to the user through the "Detection" tab. We recall that the anomaly detection procedure described in Chapter 5 consists of three steps, which are a detection of temporal change-points, a spatial clustering of the stations and a procedure for anomaly detection on a given time segment. The results of the first two steps are precomputed and stored in the input file (see Section 6.1), whereas the third step is performed interactively in the application.

Visualisation of temporal segmentation results

The temporal segmentation results are spread across two panels. The first one displays time segmentations performed on the daily maximum time series assuming a Weibull distribution model. Several elements are made available to the user:

- The costs of different segmentations are plotted against their number of change points. The user can choose a penalty value with a slider (see Figure 6.4). For the selected penalty value, the corresponding segmentation is highlighted in red. The red lines on the graph are the estimates of a piece-wise linear model with two components obtained using the elbow method Algorithm 2. The vertical black line indicates the optimal number of change-points detected by CROPS as the abscissa of the elbow position location. The application is initialized on the penalty value corresponding to this optimal number of change-points, however the user may select a non optimal segmentation, which we did on purpose in the figures (see Figure 6.4).

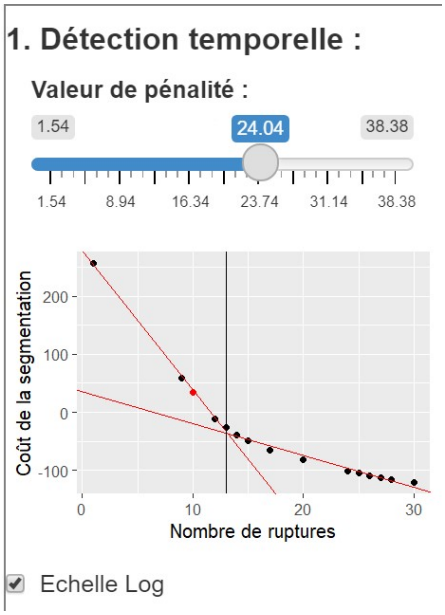

Figure 6.4: Penalty choice and corresponding segmentation information.

- A plot of the segmentation of the daily maximum concentration signal (see Figure 6.5). This plot is interactive: it is automatically updated when the user moves the cursor to

change the penalty value, and segments can also be selected with a mouse-click. In this case, it is highlighted in black (see Figure 6.5 for an illustration). To better distinguish quantified low concentration values from values under the quantification threshold (which are censored to the current LOQ value), we added the possibility to plot the time series in logarithmic scale to obtain a better visualization. In Figure 6.5 concentrations are plotted in logarithmic scale.

Figure 6.5: Plot of the resulting segmentation.

A second panel is available for the user, which content is dependent on the segment selected in the first panel. In this second panel, information specific to the segment selected by the mouse-click feature is displayed:

- The dates corresponding to the segment temporal limits, the number of daily maximum concentration values inside the segment; the percentage of quantified concentration measurements within the segment; the number of active stations within the segment timespan as well as the minimum, mean, median and maximum values of daily maximum concentrations.

- A plot displaying the parametric cumulative distribution function computed with the estimated parameter values and the empirical one, for visual comparison (see Figure 6.6 left). On this plot, vertical black lines are located at the LOQ values. The empirical cdf is coloured in blue and the one obtained with the parametric model is coloured in red. This plot aims at providing a first visual evaluation of the relevance of the parametric distribution used to model the data on the selected segment (a Weibull distribution in the current implementation). The illustrative example in Figure 6.6 confirms the choice of the Weibull distribution as a relevant model for Prosulfocarb concentration data in surface waters.

- A seasonal plot, providing a comparison between the distribution of the data in the selected time segment and the distribution of the data the same time periods on the previous

and following years. The comparison is done through violin plots. For instance, in Figure 6.6-right, the selected segment spans from January 22nd 2018 to October 5th 2018. We represented the violin plot of the daily maximum concentrations between January 22nd and October 5th of each year available. The violin plot of the selected segment is highlighted in red. Note that the violin plots adapt to whether the logarithmic scale was chosen or not in the previous panel.

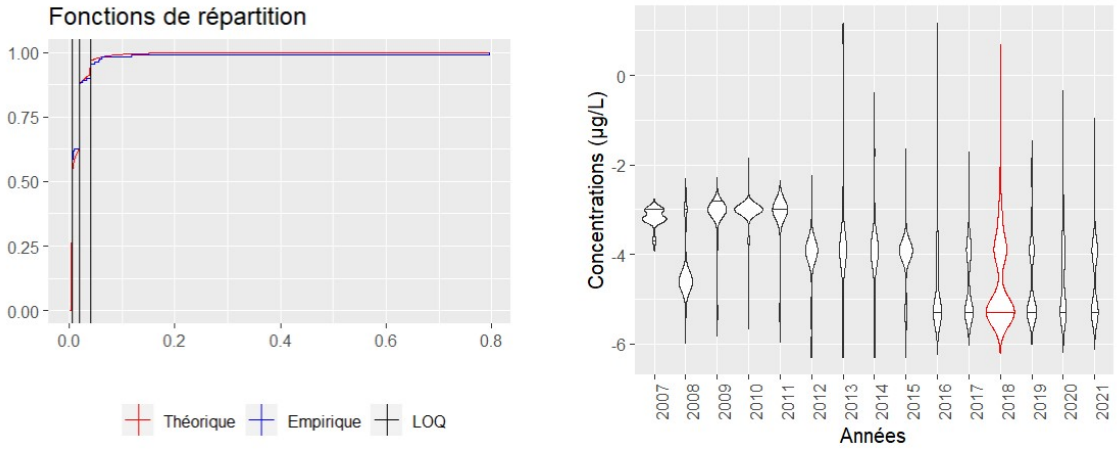

Figure 6.6: Informations on the selected segment.

These two panel encompass the whole temporal detection procedure. The spatial clustering and anomaly detection are presented in the next section. Following the monitoring procedure developed in Chapter 5, the spatial clustering is independent of the segmentation of the daily maximum series. However, anomaly detection is made conditionally to the segment selected by the user. All anomaly detection information shown in the following figures is then dependent on the segment highlighted in black in Figure 6.5.

Spatial clustering and anomaly detection

As stated previously, the information provided to the user in this part of the application is dependent of the temporal segment selected in the temporal segmentation panel (see section 6.2.2). The spatial clustering and the anomaly detection steps are covered by two additional panels.
The first one displays a map of the geographical area under consideration and a drop-down menu (see Figure 6.7) allowing to choose the information to be displayed on the map. Four different options are available in the menu:

- "Active station": the stations are plotted with two different colors according to whether they collected a sample during the segment timespan or not.
- "Hydrographic component": the stations are plotted and coloured according to the connected component of the station graph they belong to; note that this information is already available in the "Home tab" but it was repeated here for the user convenience.

- "Spatial clustering": the stations are plotted and coloured according to the cluster they belong to (see Figure 6.8).
- "Pareto front values": stations are plotted and coloured according to their Pareto front level (see Figure 6.9).

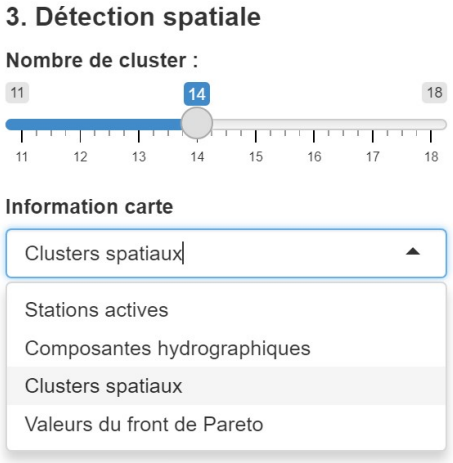

Figure 6.7: Drop down menu with a slider selecting the number of clusters in the spatial clustering.

When "Spatial clustering" or "Pareto front values" are selected, it is necessary to select the desired number of clusters. This number can be selected with a cursor located on the left of the map (see Figure 6.7). The map updates automatically with the selected number of clusters.

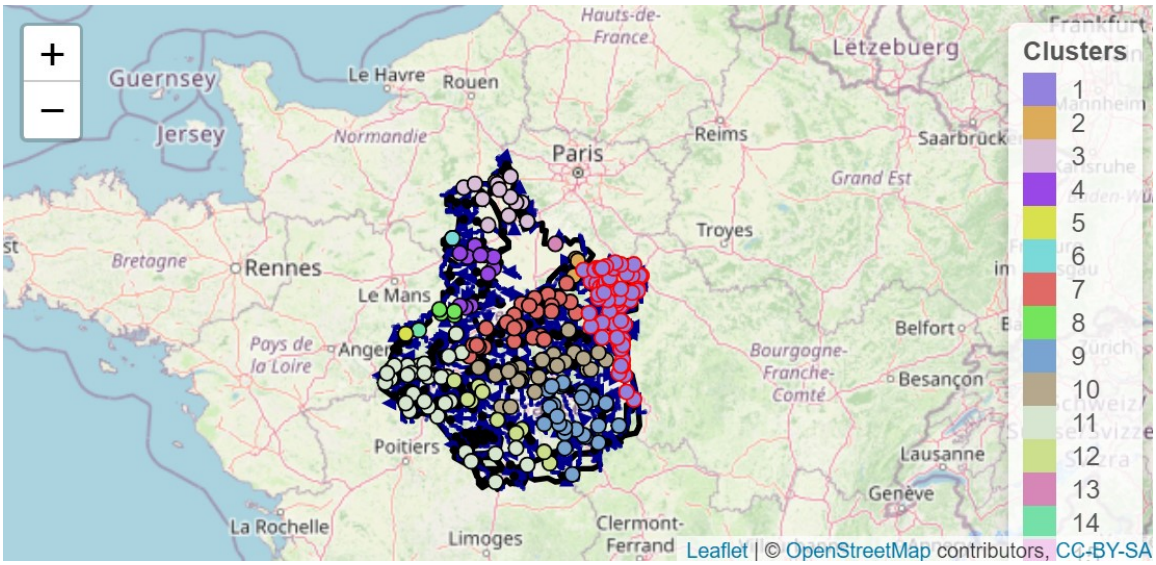

Figure 6.8: Map displaying the clusters. The clustering selected is composed of 14 clusters.

For "Spatial clustering", the information displayed was computed during the preprocessing step and stored in the input file of the application.

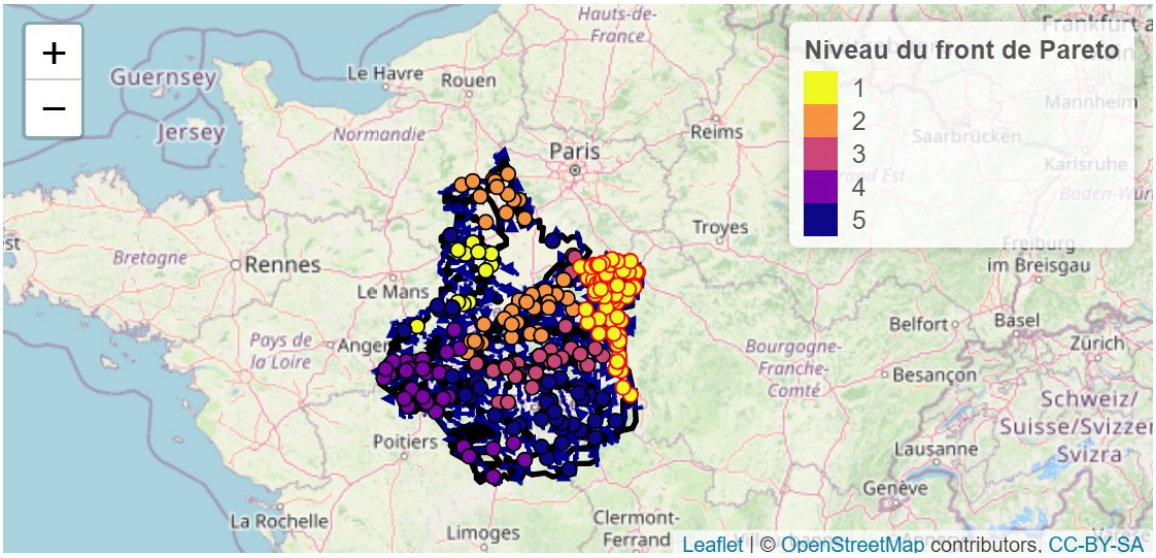

Figure 6.9: Map displaying the Pareto front values of each cluster.

For "Pareto front values", the Pareto level values are computed and refreshed upon each change in the number of cluster using the `Rpref` package. The `Rpref` package implements the multi-criteria optimization using the skyline algorithm (Borzsony et al., 2001). Additionally, a cluster selection feature is implemented: upon selection of a station, the cluster in which the station is located is highlighted in red as in Figure 6.9.
The last panel is dependent on all choices made in the previous panels. It is composed of three windows:

- A Pareto plot where clusters are plotted according to the criteria $\bar{W}_m$ $\bar{I}_m$ described in Section 5.2.3 (see Figure 6.10). The selected cluster in the map is also highlighted in red in the plot.

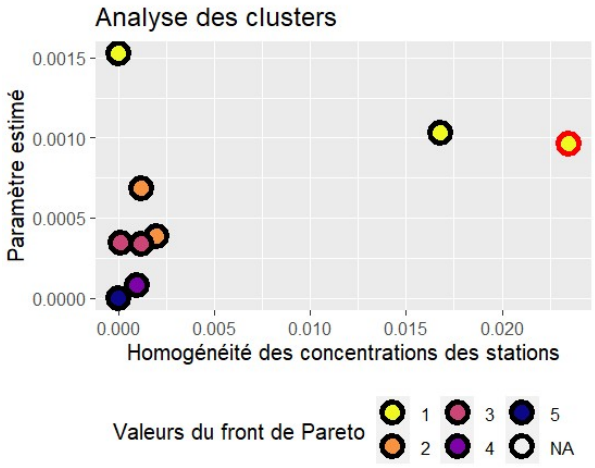

Figure 6.10: Plot of the Pareto front.

- Additional text information is provided on all measurements in the selected cluster, such as:
  - The total number of measurements collected.
  - The percentage of quantified concentration results among these measurements.
  - The number of active stations in the cluster.
  - The minimum, the mean, the median and the maximum of concentrations.
  - The different LOQ values present in the measurements (the most frequent one being indicated).
  - The station ID with the highest quantification rate and its associated number of measurements and the quantification rate information.
- A station measurement plot (see Figure 6.11): once a station has been selected in the map, this plot displays its measurements during the selected time segment.

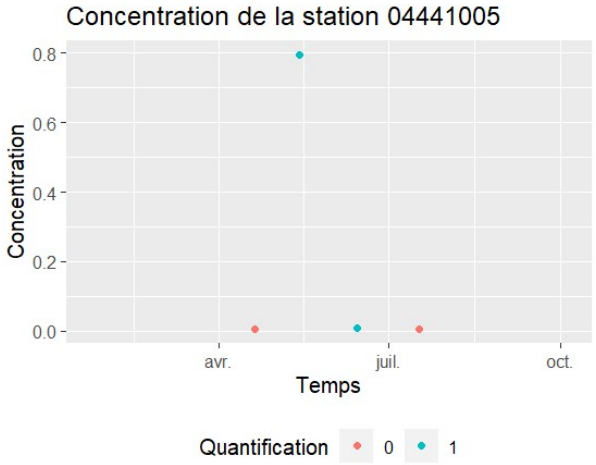

Figure 6.11: Selected station sample values during the selected temporal segment.

The aim of the application is to alert the expert to some clusters with high values of $\bar{W}_m$ and $\bar{I}_m$. The application allows the expert to quickly identify geographical areas and time periods where unusual concentration values have occurred. However, the actual abnormality of the detected signals must be assessed by the experts.

7. Conclusion and perspectives

This thesis proposes an original method to extract spatio-temporal information from environmental data. We saw in Chapter 2 all the challenges that the ANSES is faced with. Time sampling heterogeneity and spatio-temporal heterogeneity together with specific censoring issues for concentration data demand the construction of sophisticated modelling and estimation methods and the design of relevant indicators to help experts in their monitoring tasks. A user-friendly implementation of these outcomes is also needed.
Chapters 3, 4 and 5 cover the full description of the new method and data exploration process developed to address the above-mentioned issues. Our approach is first driven by temporal detection. We aim at identifying time periods where the observed environmental concentrations are homogeneously distributed. Chapter 3 reviewed a selection of of change-point detection methods that seemed well adapted in our context,
In Chapter 4, we focused both on the effects of censoring on change-point detection methods, for models where some parameters are specific to each segment while others are shared by all segments. We derived an adapted optimisation method for our modelling framework. Chapter 5 provided the principles underlying our spatio-temporal detection method through a practical implementation on a real-life dataset, namely the concentration of Prosulfocarb in surface waters in the Centre-Val de Loire region between January 9, 2007 and April 8, 2022.
We aggregated the temporal information at the regional scale to form series of daily maximum concentrations. The change point detection method proposed in Chapter 4 was then used on this series to reveal temporal regimes that proved to be good indicators of changes in farming activity. In parallel, the stations monitoring surface water quality were modelled according to a graph. In this graph, the edges between the stations and their associated weights were computed from the spatial information on the river system in the region. The graph was then clustered to derive groups of stations, the resulting clusters constitute the aggregated spatial information. Finally, for each time regime, we used the available temporal and spatial information to propose a comparison of these clusters based on multi-criteria optimisation and identify the most anomalous ones. We hope that this detection constitutes a first useful indicator for experts in ANSES.
Chapter 6 is the practical implementation of this procedure under the form of an `R` script for a preprocessing and precomputation step feeding an interactive Rshiny application. The design of this application results from discussions with the experts of the ANSES.
This thesis demonstrates that it is possible to extract information from spatio-temporal environmental data whose characteristics make modelling difficult. Some natural future developments emerge from this work.
A first axis of investigation concerns developments aiming to improve the modelling and analysis of concentration data. We could further advance the analysis of pesticide concentrations by introducing a multivariate modelling framework in order to take simultaneous substances into account. Such methods are introduced in Pickering (2016). This would broaden the scope of environmental monitoring to substance associations monitoring. The comparison of temporal

change-points positions in different substances concentrations is possible with the works of Cleynen & Robin (2014). Observing similar change-points positions in different substances would imply a strong association in their use. The evolution in time of the spatial distribution of anomalous clusters can also be another crucial point. We showed in Chapter 5 that the spatial distribution of Pareto levels did not seem to be uniform. Analysing the time series of the cluster Pareto levels could uncover additional information.

Irregular sampling, both in time and space, is another major issue in environmental data, therefore another axis could consist in improving the sampling procedure. There is currently a high spatio-temporal heterogeneity in the collected concentration data, which is damped through time and space aggregation in the analyses. This irregular sampling prevents from observing the dispersion dynamics of a substance in space and time. Providing a strategy for improving station locations and sampling schedules can be seen as an optimal design problem (Müller et al., 2011; Marsh & Ewers, 2012). This could prove to be a harsh task because optimal design problems would be highly constrained by the spatial structure underlying the station locations. For instance, the constraints on the location of stations monitoring air or surface waters quality are drastically different. The first case allows for almost any position in a given area, the latter is more constrained: stations have to be located on riverside (or at least near a surface water body).

Finally, from the implementation point of view, more work would be needed to go from the current prototype to a software allowing to handle larger datasets at the French national scale, and ultimately be able to compute and display indicators for several substances and any sub-regions. Integrating additional information such as the sales of the substance loaded in the application and the spatial distribution of the crops targeted by the substance would help in the monitoring mission.

References

Agriculture & Environment Research Unit (AERU) at the University of Hertfordshire. (n.d.). *Prosulfocarb (Ref: SC 0574).* `https://sitem.herts.ac.uk/aeru/ppdb/en/Reports/557.htm`. Retrieved from `http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/557.htm` (Retrieved: March 1, 2022. Part of Lewis et al. (2016))

Andrienko, N., Andrienko, G., & Gatalsky, P. (2003). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing*. doi: 10.1016/s1045-926x(03)00046-6

Ansari, M. Y., Ahmad, A., Khan, S. S., Bhushan, G., & Mainuddin. (2019). Spatiotemporal clustering: a review. *Artificial Intelligence Review*. doi: 10.1007/s10462-019-09736-1

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

Arlot, S., & Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research*. doi: 10.1145/1577069.1577079

Auger, I. E., & Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*. doi: https://doi.org/10.1016/S0092-8240(89)80047-3

Aznar, R., Moreno-Ramón, H., Albero, B., Sánchez-Brunete, C., & Tadeo, J. L. (2017). Spatiotemporal distribution of pyrethroids in soil in mediterranean paddy fields. *Journal of Soils and Sediments*. doi: 10.1007/s11368-016-1417-2

Bai, J. (1994). LEAST SQUARES ESTIMATION OF a SHIFT IN LINEAR PROCESSES. *Journal of Time Series Analysis*. doi: 10.1111/j.1467-9892.1994.tb00204.x

Baran, N., Rosenbom, A. E., Kozel, R., & Lapworth, D. (2022). Pesticides and their metabolites in european groundwater: Comparing regulations and approaches to monitoring in france, denmark, england and switzerland. *Science of The Total Environment*. doi: 10.1016/j.scitotenv.2022.156696

Bardet, J.-M., Kengne, W., & Wintenberger, O. (2012). Multiple breaks detection in general causal time series using penalized quasi-likelihood. *Electronic Journal of Statistics*. doi: 10.1214/12-ejs680

Bardet, Jean-Marc, Brault, Vincent, Dachian, Serguei, Enikeeva, Farida, & Saussereau, Bruno. (2020). Change-point detection, segmentation, and related topics. *ESAIM: ProcS*. doi: 10.1051/proc/202068006

Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt change: Theory and application*. prentice Hall Englewood Cliffs.

Baudry, J.-P., Maugis, C., & Michel, B. (2011). Slope heuristics: overview and implementation. *Statistics and Computing*. doi: 10.1007/s11222-011-9236-1

Birgé, L., & Massart, P. (2006). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*. doi: 10.1007/s00440-006-0011-8

Borzsony, S., Kossmann, D., & Stocker, K. (2001). The skyline operator. In *Proceedings 17th international conference on data engineering* (p. 421-430).

Bouveyron, C., Jacques, J., Schmutz, A., Simoes, F., & Bottini, S. (2022). Co-clustering of multivariate functional data for the analysis of air pollution in the south of france. *Annals of Applied Statistics*. doi: 10.1214/21-AOAS1547

Bunce, C., Carr, J. R., Nienow, P. W., Ross, N., & Killick, R. (2018). Ice front change of marine-terminating outlet glaciers in northwest and southeast greenland during the 21st century. *Journal of Glaciology*. doi: 10.1017/jog.2018.44

Ccanccapa, A., Masiá, A., Andreu, V., & Picó, Y. (2016). Spatio-temporal patterns of pesticide residues in the turia and júcar rivers (spain). *Science of The Total Environment*. doi: https://doi.org/10.1016/j.scitotenv.2015.06.063

Chapter 23 nonparametric tests for trend detection. (1994). In K. W. Hipel & A. I. McLeod (Eds.), *Time series modelling of water resources and environmental systems.* doi: https://doi.org/10.1016/S0167-5648(08)70688-9

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

Chen, J., Kim, S.-H., & Xie, Y. (2020). S3t: A score statistic for spatiotemporal change point detection. *Sequential Analysis*. doi: 10.1080/07474946.2020.1826796

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.*

Cleynen, A., & Robin, S. (2014). Comparing change-point location in independent series. *Statistics and Computing*. doi: 10.1007/s11222-014-9492-y

Cohen, A. C. (1965). Maximum likelihood estimation in the weibull distribution based on complete and on censored samples. *Technometrics*. doi: 10.1080/00401706.1965.10490300

Costa, M., Gonçalves, A. M., & Teixeira, L. (2016). Change-point detection in environmental time series based on the informational approach. *Electronic Journal of Applied Statistical Analysis*. doi: 10.1285/i20705948v9n2p267

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

Croghan, W., & 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.*

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*. Retrieved from `https://igraph.org`

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An r package for fitting distributions. *Journal of statistical software*. doi: 10.18637/jss.v064.i04

de Solla, S. R., Weseloh, D. C., Hebert, C. E., & Pekarik, C. (2010). Impact of changes in analytical techniques for the measurement of polychlorinated biphenyls and organochlorine pesticides on temporal trends in herring gull eggs. *Environmental Toxicology and Chemistry*. doi: 10.1002/etc.191

Einmahl, J. H. J., & McKeague, I. W. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*. doi: 10.3150/bj/1068128978

Endres, M., Roocks, P., & Kießling, W. (2015). Scalagon: an efficient skyline algorithm for all seasons. In *International conference on database systems for advanced applications.*

Faure, C., Bardet, J.-M., Olteanu, M., & Lacaille, J. (2016). Comparison of three algorithms for parametric change-point detection. In *Esann.*

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*. doi: 10.1016/j.patrec.2005.10.010

Fearnhead, P., Maidstone, R., & Letchford, A. (2018). Detecting changes in slope with an l_0 penalty. *Journal of Computational and Graphical Statistics*. doi: 10.1080/10618600.2018.1512868

Fearnhead, P., & Rigaill, G. (2018). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*. doi: 10.1080/01621459.2017.1385466

Figueiredo, D. M., Duyzer, J., Huss, A., Krop, E. J., Gerritsen-Ebben, M., Gooijer, Y., & Vermeulen, R. C. (2021). Spatio-temporal variation of outdoor and indoor pesticide air concentrations in homes near agricultural fields. *Atmospheric Environment*. doi: https://doi.org/10.1016/j.atmosenv.2021.118612

Fomby, T. B., & Lin, L. (2006). A change point analysis of the impact of "environmental federalism" on aggregate air quality in the united states: 1940-98. *Economic Inquiry*. doi: 10.1093/ei/cbj006

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*. doi: 10.1016/j.physrep .2009.11.002

Frick, K., Munk, A., & Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. doi: https://doi.org/10.48550/ arXiv.1301.7212

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*. doi: 10.1214/14-aos1245

Gillaizeau, F., Gal, C. L., Maudet, C., Fournier, M., & Leuillet, S. (2020). Méthodes de gestion des valeurs sous des seuils de détection ou de quantification. *Revue d'Épidémiologie et de Santé Publique*. doi: 10.1016/j.respe.2020.03.076

Harchaoui, Z., Moulines, E., & Bach, F. (2008). Kernel change-point analysis. In *Advances in neural information processing systems.*

Haynes, K., Eckley, I. A., & Fearnhead, P. (2017). Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*. doi: 10.1080/10618600.2015.1116445

Haynes, K., Fearnhead, P., & Eckley, I. A. (2016). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*. doi: 10.1007/s11222-016-9687 -5

He, H., & Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*. doi: 10.3150/09-bej232

Hébrail, G., Hugueney, B., Lechevallier, Y., & Rossi, F. (2010). Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*. doi: https:// doi.org/10.1016/j.neucom.2009.11.022

Höhle, M. (2010). Online change-point detection in categorical time series. In *Statistical modelling and regression structures.* doi: https://doi.org/10.1007/978-3-7908-2413-1_20

Institut National de l'Information Géographique et Forestière. (n.d.). *BD TOPO®*. `https://geoservices.ign.fr/bdtopo`. Retrieved from `https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo` (Retrieved: March 1, 2022)

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions.* Wiley-Interscience.

Jørgensen, L. F., & Stockmarr, J. (2008). Groundwater monitoring in denmark: characteristics, perspectives and comparison with other countries. *Hydrogeology Journal*. doi: 10.1007/ s10040-008-0398-7

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

Kießling, W. (2002). Chapter 28 - foundations of preferences in database systems. In *Vldb '02: Proceedings of the 28th international conference on very large databases.* doi: https://doi.org/10.1016/B978-155860869-6/50035-4

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*. doi: 10.1080/01621459.2012.737745

Ko, D. R., Chung, S. P., You, J. S., Cho, S., Park, Y., Chun, B., ... Hong, J. H. (2017). Effects of paraquat ban on herbicide poisoning-related mortality. *Yonsei Medical Journal*. doi: 10.3349/ymj.2017.58.4.859

Laroche, C., Olteanu, M., & Rossi, F. (2021). Estimation paramétrique de ruptures dans des données censurées à gauche. In *52ème journées de statistique de la société française de statistique (sfds).*

Laroche, C., Olteanu, M., & Rossi, F. (2022a). Pesticide concentration monitoring: Investigating spatio-temporal patterns in left censored data. *Environmetrics*. doi: 10.1002/env.2756

Laroche, C., Olteanu, M., & Rossi, F. (2022b). Pesticide concentration monitoring: investigating spatio-temporal patterns in left censored data. In *15-th international conferenceon operation research.*

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*. doi: https://doi.org/10.1016/S0304-4149(99)00023-X

Lavielle, M., & Lebarbier, E. (2001). An application of MCMC methods for the multiple change-points problem. *Signal Processing*. doi: 10.1016/s0165-1684(00)00189-4

Lévy-Leduc, C., & Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*. doi: 10.1214/08-AOAS232

Lewis, K., Tzilivakis, J., Warner, D., & Green, A. (2016). An international database for pesticide risk assessments and management. *Human and Ecological Risk Assessment: An International Journal*. doi: 10.1080/10807039.2015.1133242

Li, S., Xie, Y., Dai, H., & Song, L. (2015). M-statistic for kernel change-point detection. In *Advances in neural information processing systems.*

Li, W., Guo, W., Luo, X., & Li, X. (2010). On sliding window based change point detection for hybrid SIP DoS attack. In *2010 IEEE asia-pacific services computing conference.*

Li, Y., Bao, T., Shu, X., Gao, Z., Gong, J., & 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*. doi: 10.1007/s13349-021-00520-w

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., & Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and ecological statistics*. doi: 10.1007/s10651-013-0261-4

Liu, C., Chen, Y., Chen, F., Zhu, P., & Chen, L. (2022). Sliding window change point detection based dynamic network model inference framework for airport ground service process. *Knowledge-Based Systems*. doi: 10.1016/j.knosys.2021.107701

Liu, S., Wright, A., & Hauskrecht, M. (2017). Change-point detection method for clinical decision support system rule monitoring. In *Conference on artificial intelligence in medicine in europe.*

Lung-Yut-Fong, A., Lévy-Leduc, C., & Cappé, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*. Retrieved from `http://www.numdam.org/item/JSFS_2015__156_4_133_0/`

Maidstone, R. (2016). *Efficient analysis of complex changepoint problems*. Lancaster University (United Kingdom). Retrieved from `https://eprints.lancs.ac.uk/id/eprint/83055/1/2016MaidstonePhD.pdf`

Maimon, O., & Rokach, L. (2010). *Data mining and knowledge discovery handbook*. Springer US.

Majumdar, A., Gelfand, A. E., & Banerjee, S. (2005). Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference*. doi: https://doi.org/10.1016/j.jspi.2003.08.022

Manly, B. F. (2008). *Statistics for environmental science and management*. Chapman and Hall/CRC.

Marchant, C., Leiva, V., Christakos, G., & Cavieres, M. F. (2018). Monitoring urban environmental pollution by bivariate control charts: New methodology and case study in santiago, chile. *Environmetrics*. doi: 10.1002/env.2551

Marsh, C. J., & Ewers, R. M. (2012). A fractal-based sampling design for ecological surveys quantifying β-diversity. *Methods in Ecology and Evolution*. doi: 10.1111/j.2041-210x.2012.00256.x

Menger, J. P., Ribeiro, A. V., Potter, B. D., & Koch, R. L. (2022). Change-point analysis of lambda-cyhalothrin efficacy against soybean aphid (aphis glycines matsumura): identifying practical resistance from field efficacy trials. *Pest Management Science*. doi: 10.1002/ps.7006

Müller, W. G., Rodríguez-Díaz, J. M., & López, M. J. R. (2011). Optimal design for detecting dependencies with an application in spatial ecology. *Environmetrics*. doi: 10.1002/env.1132

National Center for Biotechnology Information. (n.d.). *PubChem Compound Summary for CID 62020, Prosulfocarb.* https://pubchem.ncbi.nlm.nih.gov/compound/Prosulfocarb. Retrieved from https://pubchem.ncbi.nlm.nih.gov/compound/Prosulfocarb (Retrieved: March 1, 2022)

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*. doi: https://doi.org/10.1137/S003614450342480

Nougadère, A., Merlo, M., Héraud, F., Réty, J., Truchot, E., Vial, G., ... 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*. doi: https://doi.org/10.1016/j.foodcont.2013.12.025

Novic, A. J., O'Brien, D. S., Kaserzon, S. L., Hawker, D. W., Lewis, S. E., & Mueller, J. F. (2017). Monitoring herbicide concentrations and loads during a flood event: A comparison of grab sampling with passive sampling. *Environmental Science and Technology*. doi: 10.1021/acs.est.6b02858

Office français de la biodiversité. (n.d.). *Naïades, données sur la qualité des eaux de surface.* http://www.naiades.eaufrance.fr/, http://www.ofb.gouv.fr/. Retrieved from https://naiades.eaufrance.fr/ (Retrieved: March 1, 2022)

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

Perron, P., et al. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*. Retrieved from https://www.eco.uc3m.es

Pettitt, A. (1980). Some results on estimating a change-point using non-parametric type statistics. *Journal of Statistical Computation and Simulation*. doi: 10.1080/00949658008810413

Pickering, B. J. (2016). *Changepoint detection for acoustic sensing signals*. Lancaster University (United Kingdom). Retrieved from https://eprints.lancs.ac.uk/id/eprint/81171/1/2016PickeringPhd.pdf

Pohlert, T. (2020). trend: Non-parametric trend tests and change-point detection [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=trend (R package version 1.1.4)

Ranganathan, A. (2010). Pliss: Detecting and labeling places using online change-point detection. *Robotics: Science and Systems VI*. doi: 10.1007/s10514-012-9273-4

Reeves, J., Chen, J., Wang, X. L., Lund, R., & Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology*. doi: https://doi.org/10.1175/JAM2493.1

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to $k\_\{max\}$ change-points. *Journal de la Société Française de Statistique*. doi: https://doi.org/10.48550/arXiv.1004.0887

Roocks, P. (2016). Computing Pareto Frontiers and Database Preferences with the rPref Package. *The R Journal*. doi: 10.32614/RJ-2016-054

Ryberg, K. R., Hodgkins, G. A., & Dudley, R. W. (2020). Change points in annual peak streamflows: Method comparisons and historical change points in the united states. *Journal of Hydrology*. doi: 10.1016/j.jhydrol.2019.124307

Sadani, S., Abdollahnezhad, K., Teimouri, M., & Ranjbar, V. (2019). A new estimator for weibull distribution parameters: Comprehensive comparative study for weibull distribution. *arXiv preprint arXiv:1902.05658*. doi: https://doi.org/10.52547/jsri.16.1.33

Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., & Kaufman, J. D. (2011). Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*. doi: https://doi.org/10.1016/j.atmosenv.2011.04.073

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*. doi: https://doi.org/10.1016/j.cosrev.2007.05.001

Shi, X., Beaulieu, C., Killick, R., & Lund, R. (2022). Changepoint detection: An analysis of the central england temperature series. *Journal of Climate*. doi: 10.1175/JCLI-D-21-0489.1

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., & 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*. doi: 10.1109/MSP.2012.2235192

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*. doi: https://doi.org/10.1016/j.sigpro.2019.107299

Villani, C. (2009). *Optimal transport: old and new*. Springer.

Wang, Y., Wang, Z., & Zi, X. (2019, apr). Rank-based multiple change-point detection. *Communications in Statistics - Theory and Methods*, *49*(14), 3438–3454. doi: 10.1080/03610926.2019.1589515

Wasson, J., Chandesris, A., Pella, H., & 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* (Tech. Rep.). irstea. Retrieved from `https://hal.inrae.fr/hal-02580774`

Web resource: Ades portal. (n.d.). `https://ades.eaufrance.fr/Recherche`.

Web resource: Agreste. (n.d.). `https://agreste.agriculture.gouv.fr/agreste-web/download/methode/S-PK%20Viticulture%202019/20191217_questionnaire_PK_Viti.pdf`.

Web resource: Agreste source. (n.d.). https://agreste.agriculture.gouv.fr/agreste-web/disaron/Chd2009/detail/.

Web resource: Anses decision site. (n.d.). https://www.anses.fr/fr/decisions.

Web resource: Anses laboratories mandates. (n.d.). https://www.anses.fr/fr/system/files/ANSES-Ft-PlaquetteMandats_FR.pdf.

Web resource: European commission. (n.d.). https://ec.europa.eu/info/sites/default/files/research_and_innovation/funding/documents/ec_rtd_he-partnerships-chemical-risk-assessment.pdf.

Web resource: Geodair portal. (n.d.). https://www.geodair.fr/donnees/consultation.

Web resource: Météo-france data (SYNOP). (n.d.). https://public.opendatasoft.com/explore/dataset/donnees-synop-essentielles-omm/table/?flg=fr&sort=date.

Web resource: Naiades portal. (n.d.). https://naiades.eaufrance.fr/acces-donnees#/physicochimie.

Web resource: AGRITOX database. (n.d.). https://www.data.gouv.fr/fr/datasets/base-de-donnees-agritox/.

Web resource: BDTOPO database. (n.d.). https://geoservices.ign.fr/bdtopo.

Web resource: IGN data. (n.d.). https://geoservices.ign.fr/rpg.

Yang, T. Y., & Kuo, L. (2001). Bayesian binary segmentation procedure for a poisson process with multiple changepoints. *Journal of Computational and Graphical Statistics*. doi: 10.1198/106186001317243449

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz' criterion. *Statistics and Probability Letters*. doi: https://doi.org/10.1016/0167-7152(88)90118-6

Zhang, Q., Li, Z., Zeng, G., Li, J., Fang, Y., Yuan, Q., ... Ye, F. (2008). Assessment of surface water quality using multivariate statistical techniques in red soil hilly region: a case study of xiangjiang watershed, china. *Environmental Monitoring and Assessment*. doi: 10.1007/s10661-008-0301-y

Zou, C., Yin, G., Feng, L., & Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*. doi: 10.1214/14-aos1210

Appendices

A. Chapter 4 supplementary material

This appendix chapter discusses various details presented in Chapter 4. In Section A.1, we prove that the estimators $\widehat{K}$ and $\widehat{\mathcal{T}}$ still converge in the censored setting using elements of the demonstration made in Lavielle (1999). Then, we study the convergence of a segment parameters and the importance of the initialization value in the Newton-Raphson method in Section A.2. Lastly in Section A.3, we check if the necessary conditions to use the PELT algorithm are verified.

A.1. Elements of proof of convergence of the parametric change-point detection model

The proof of convergence of (4.5) is based on the approach developed in Lavielle (1999). The most critical element in the proof is the condition C0(h) of Lavielle (1999). Let's denote $\eta_i = \ln f(Y_i, \theta_k) - \mathbb{E}[\ln f(Y_i, \theta_k)]$ for i belonging to the k-th segment and associated to the parameters θ_k. We have the following proposition:

Proposition A.1.1. *There exists $C < \infty$ such that for any $t \geq 0$ and any $s > 0$,*

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

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

In Lavielle (1999), an indication is given 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 is also a sequence of random variables and the proposition is verified for $h = 1$. Then from Theorem 2.2 of Lavielle (1999), we have the consistency of the estimator:

$$(\widehat{\mathcal{T}}, \widehat{\boldsymbol{\theta}}) \xrightarrow[n\to\infty]{\mathbb{P}^\star} (\boldsymbol{\mathcal{T}}^\star, \boldsymbol{\theta}^\star)$$

$(Y_1, ..., Y_n)$ is not i.i.d., the censoring threshold makes the support of this random variables differ. However, the base signal is emanating from $(C_1, ..., C_n)$ that is defined in 4.1 and that is i.i.d.. We can then suppose that $C0(h)$ is verified.

A.2. Newton-Raphson initialization experiments

We show in this section that the initialization value is an important parameter when estimating the segment parameters with numerical methods such as Newton-Raphson. We illustrate this fact with Q set as the Weibull distribution.

We propose to test out four initialisation values in the Newton-Raphson algorithm. We can choose between the classical techniques such as the moment method estimator λ_{init}^{MM} Johnson et al. (1994), the quantile inversion estimator λ_{init}^{QI}, the weighed maximum likelihood estimator λ_{init}^{WMLE} Sadani et al. (2019) 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 :

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

where $q_{\mathcal{W}(n,n)}^{0.5}$ is the median of Weibull with parameters (n, n), $q_{\boldsymbol{x}}^{\epsilon}$ is the ϵ-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 section 4.3). Second, these estimators do not take the censoring 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 scale 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 A.1 and A.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.

A.3. Verifying PELT assumptions

This Section uses the same notation than Section 3.4. Some necessary conditions must be met before using the PELT algorithm. It can be found in Theorem 3.1 of Killick et al. (2012) and

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 A.1: Choice of initialization value: simulation results for $n = 20$.

can be stated as follow:

Proposition A.3.1. *We assume that when introducing a change point 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 $t < s < T$,*

$$W(y_{t:s}) + W(y_{s:T}) + K \leq W(y_{t:T}) \tag{A.2}$$

Then if

$$F(t) + W(y_{t:s}) + K \geq F(s) \tag{A.3}$$

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

Proof: The equation A.2 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.3 is true. Adding $W(y_{s:T}, \hat{\lambda}_{s:T})$ on both sides of the inequality 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.

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 A.2: Choice of initialization value: simulation results for $n = 100$.

B. Chapter 6 supplementary material

B.1. Clustering selected for the application

In order to choose the clusterings that are imported in to the application, we use the elbow method described in Algorithm 2. Instead of plotting the clustering inertia values against their number of clusters, we use the logarithm of the inertia. The decrease of the inertia don't seem to have any linear behaviour. Two elbows appeared in the decrease of the logarithm though.

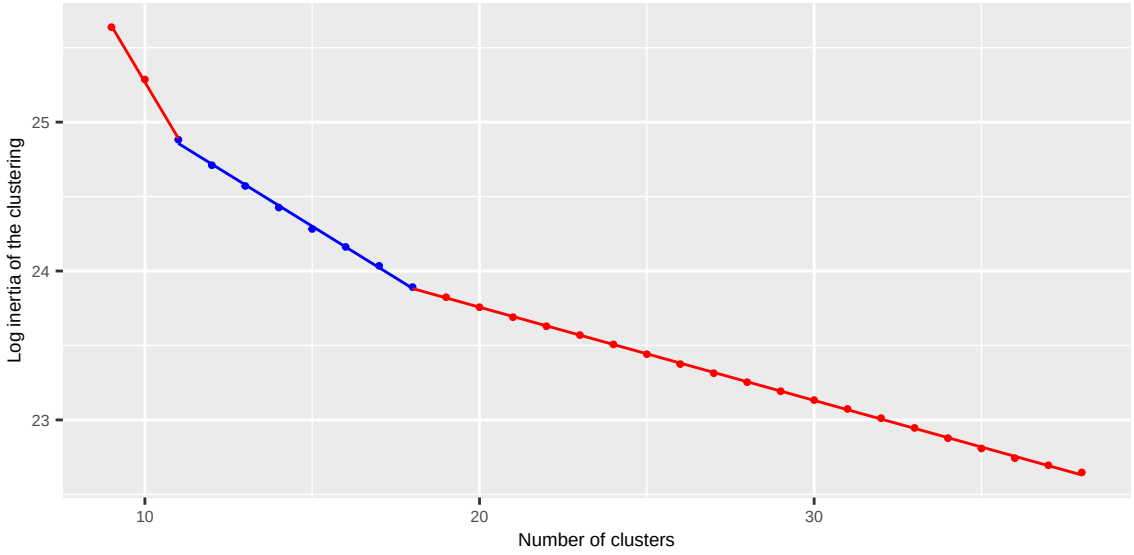

Figure B.1: Clustering candidates selected in the application.

Instead of fitting a bipartite model, we fitted the best piece-wise linear model composed of three regression models as represented in Figure 5.9. All the clustering that had their log-inertia located on the middle regression line were selected. They are coloured in blue in Figure 5.9.

B.2. Application explanatory note

Notice

Clément LAROCHE : clement-laroche@hotmail.fr

Dernière mise à jour : 29/03/2022

On présente dans cet onglet la notice d'utilisation de l'application. On rappelle que cette application a pour objectif d'aider l'utilisateur à localiser des signaux anormaux de concentration dans le temps et dans le territoire. Cet outil *n'a pas vocation ou prétention* d'expliquer la présence de phénomènes anormaux. Cela relève de l'expertise de l'utilisateur.

Onglet présentation

Le premier onglet donne des informations générales sur les données qui ont été chargées dans l'application.

Encadré 1 : Informations générales

On introduit la définition suivante :

Définition 1 (maximum journalier) : En se fixant une date d, on définit un maximum journalier comme le maximum des relevés de concentrations ayant eu lieu dans la région d'étude le jour d. L'information de quantification du maximum est également conservée :

- si le maximum journalier correspond à une valeur de LOQ alors il est dénoté comme non quantifié.
- si le maximum journalier ne correspond pas à une valeur de LOQ alors il est dénoté comme quantifié.

Un fois cette définition posée, on présente les informations générales sur l'étude que l'on s'apprête à mener en utilisant l'application :

- on donne le nom de la susbtance sur laquelle on travaille (on ne peut charger qu'une substance à la fois), correspond à l'entrée **Paramètre** de la requête effectuée sur http://www.naiades.eaufrance.fr/
- on donne la période temporelle d'étude, correspond aux entrées **Date de début** et **Date de fin** de la requête effectuée sur http://www.naiades.eaufrance.fr/
- on donne le nom de la région d'étude, correspond à l'entrée **Zone administrative** de la requête effectuée sur http://www.naiades.eaufrance.fr/
- on donne le nombre de prélèvements effectués dans cette zone et durant cette période, cela correspond au nombre de lignes du tableau **Analyses.csv** issues de la requête sur http://www.naiades.eaufrance.fr/
- on donne le pourcentage de prélèvements des relevés, on utilise la colonne **MnemoRqAna** du tableau **Analyses.csv** dont on compte le nombre d'occurences telle que la description indique que le résultat est au dessus du seuil de quantification
- on donne le nombre de jours au cours desquels un ou plusieurs prélèvements ont été effectués durant la période d'observation et dans la zone d'étude. Cela donne accessoirement le nombre de maximum journaliers que l'on peut calculer dans les données chargées dans l'application.

- on donne le pourcentage de quantification des maximums journaliers (cf. **Définition 1**) obtenus à partir des données.
- on affiche le nombre de stations actives (qui ont effectuées au moins un prélèvement) durant la période de temps définie et dans la zone d'étude choisie.

Encadré 2

Tous les maximums journaliers de la période d'étude et de la zone administrative choisie sont tracés en fonction du temps. Les points sont colorés selon leur statut de quantification (quantifié ou non quantifié). Les points rouge sont non quantifiés, les points bleu sont les maximums de concentrations quantifiés.

Encadré 3

Définition 2 (composante hydrographique) : on appelle ici composante hydrographique un réseau d'eaux de surface tel que pour n'importe quel couple de point (x, y) dans ce réseau, il existe un chemin d'eau reliant x et y. Le sens d'écoulement du courant n'est pas considéré.

On introduit la géographie de cette étude. Sont tracés sur une carte :

- toutes les stations ayant effectué au moins un prélèvement durant la période d'étude.
- en noir, les contours de la région administrative choisie pour cette étude.
- en bleu, les principaux cours d'eaux de la région (ceux qui disposent d'un code cours d'eau dans la base BD TOPO IGN)
- les stations sont coloriées selon la composante hydrographique auxquelles elles appartiennent (cf **Défintion 2**).

Toutes les coordonnées géographiques utilisées sont dans le Système de Coordonnées de Références **WGS84**.

Onglet Détection

Cet onglet constitue le principal outil qui permet de mettre en lumière des signaux anormaux présents dans les données chargées. Cette page comporte 4 encadrés. Les encadrés 1 et 2 concernent la partie détection temporelle. Les encadrés 3 et 4 sont portés sur le clustering spatial et la détection de clusters anormaux. Les choix effectués dans l'encadré 1 déterminent les informations qui seront présentes dans les encadrés suivants. L'encadré 2 est complémentaire du premier et présente des informations supplémentaires qui varient selon les choix effectués en encadré 1. Les choix de l'encadré 3 détermineront également les information dans l'encadré 4.

Encadré 1 : Détection temporelle

Cet encadré comporte 4 éléments, on commence par les deux plus intuitifs :

- un **graphique des maximums journaliers** portant le titre **Résultat de la segmentation**, les résultats de détection de ruptures temporelles sont appliqués sur la série des maximums journaliers. Les segments temporels sont représentés en blanc. Il est possible de sélectionnner un segment temporel, si c'est le cas il est surligné en noir. Choisir un segment change l'intégralité des informations de la section **Encadré 2 : Informations complémentaires sur le segment sélectionné**, de la section **Encadré 3 : Détection spatiale** et l'intégralité des informations de la section **Encadré 4 : Informations complémentaires sur la détection spatiale**
- une **tick-box échelle-log**, la représentation des maximums journaliers étant très écrasée vers les seuils de LOQ mais comportant tout de même des valeurs élevées, l'échelle logarithmique permet d'obtenir une représentation mieux répartie sur l'axe des concentrations. Ce bouton change le **graphique**

des maximums journaliers ainsi que les **violins plots saisonniers** de la section **Encadré 2 : Informations complémentaires sur le segment sélectionné**

Deux autres éléments permettent le réglage de la modélisation :

- **un curseur de valeur** permettant de régler la valeur de la pénalité. Ce curseur a une influence direct sur le **graphique des maximums journaliers**. Intuitivement, il correspond au poids que l'on associe au fait de poser une rupture dans ce graphique. Donc plus le poids est léger (plus la valeur de pénalité est faible), plus on va poser de ruptures dans le graphique. A l'inverse, plus le poids est lourd (plus la valeur de pénalité est élevée), moins on va poser de ruptures dans le graphique.

- **un graphique de coude** en dessous du **curseur de valeur**. C'est un nuage de point représentant le coût total d'une segmentation (l'ensemble des ruptures sur le graphique des maximums journaliers) en fonction du nombre de rupture. Le point rouge de ce graphique vous indique le coût de la segmentation que vous êtes en train d'explorer sur le graphique des maximums journaliers. En bougeant la valeur du **curseur**, d'autres segmentations seront explorées. Dans ce cas, le point rouge vous indiquera le coût de la segmentation que vous explorerez. On considère que la segmentation optimale (donc le nombre optimal de ruptures dans le signal) correspond au "coude" du graphique. Ce coude est indiqué ici par la droite verticale noire. Le **curseur de valeurs** commence directement à cette position (donc la segmentation optimale est affichée par défaut).

Détail : Pourquoi avoir laissé d'autres segmentations que la segmentation optimale (correspondant donc au coude) ? La démarche de cette application étant exploratoire, des segmentations autres que celle définie comme optimale permettront d'explorer des segments qui n'étaient pas présents dans la segmentation optimale et qui peuvent porter de l'information que les experts seront en mesure d'analyser.

Encadré 2 : Informations complémentaires sur le segment sélectionné

Cet encadré est entièrement dépendant de la section **Encadré 1** et notamment du segment sélectionné en noir dans le graphique des résultats de segmentation. Il est composé de trois élements que l'on peut décrire comme suit :

- le premier élement est un **texte informatif sur le segment sélectionné**. Sont affichées les informations suivantes :
  - les dates délimitant le segment sélectionné (surligné en noir) en **Encadré 1**
  - le nombre de maximum journaliers dans ce segment
  - le pourcentage de quantification de maximum journalier
  - le nombre de stations qui ont été actives entre les deux dates délimitant le segment
  - la valeur minimum des maximums journaliers dans ce segment
  - la valeur moyenne des maximums journaliers dans ce segment
  - la valeur médiane des maximums journaliers dans ce segment
  - la valeur maximum des maximums journaliers dans ce segment

- **un graphique de fonctions de répartition**, ce graphique permet de juger la qualité de l'estimation du modèle sous-jacent. En bleu, on trouve la fonction de répartition empirique (donc celles des données observées) que l'on compare avec la courbe rouge qui correspond à la courbe théorique obtenue par la modélisation. Plus ces deux courbes sont proches, plus cela signifique l'ajustement du modèle choisi est précis. Les droites noires verticales correspondent aux valeurs de LOQ présentes dans le segment sélectionné.

- **un boxplot de saisonnalité**, on trace ici le violin plot (objet dont la lecture est similaire à un box plot) des maximums journaliers de concentrations du segment selectionné dans l'**Encadré 1** en rouge. On prend la période temporelle sur laquelle s'étend le segment et on trace les violins plots des mêmes périodes pour les années précédentes et suivantes. Par exemple, si le segment sélectionné s'étend du "01-03-2017" au "24-04-2017", on trace les violin plots des "01-03-20XX" au "24-04-20XX" avec "20XX" étant les autres années disponibles dans le jeu de donnée. Si l'étendue du segment sélectionné dépasse une année, ce tracé n'est pas possible. On trace alors uniquement le violin plot du segment sélectionné.

Encadré 3 : Détection spatiale

Cet encadré est entièrement dépendant de la section **Encadré 1** et notamment du segment sélectionné en noir dans le graphique des résultats de segmentation. Il est composé de 3 éléments :

- **un menu déroulant** contenant 4 options. Ces quatre options définissent les informations disposées sur la **carte interactive leaflet**. Les quatres propositions du menu sont les suivantes :
  - **Stations actives :** indiquent les stations qui ont effectuées au moins un prélèvement pendant la période du segment sélectionné dans la section **Encadré 1** (surligné en noir). C'est une variable binaire qui indique en *TRUE* que la station a bien effectué au moins un relevé, en *FALSE* que la station n'a pas effectué de relevé.
  - **Composantes hydrographiques :** cette carte reprend la carte affiché dans l'onglet **Présentation**. Cela permet de retrouver les composantes hydrographiques connectées (cf **Définition 2**) dans la région sans avoir à changer d'onglet. Elle est tout de même dépendante du choix de segment en section **Encadré 1** car seule les stations actives durant la période de temps sont colorées selon leur appartenance à une composante hydrographique. Les stations inactives sont affichées par des points noirs plus petits.
  - **Clusters spatiaux** : cette option affiche les résultats de clustering spatial sur les stations. On regroupe les stations selon leur distance dans le réseau hydrographique. Chaque cluster est associé à une couleur. Seules les stations actives sont colorées selon leur appartenance à un cluster. Plusieurs choix de clustering sont disponibles en utilisant le **curseur de nombre de clusters** au dessus du **menu déroulant**. Les stations inactives sont affichées par des points noirs plus petits.
  - **Valeurs du front de Pareto :** cette option correspond à la représentation spatial des résultats de détection d'anomalies. Plus une valeur de front de Pareto est faible plus le cluster portant cette valeur est anormal (au sens de deux critères explicités en section **Encadré 4**). Une valeur de front de Pareto est affectée à chaque cluster, les résultats sont donc dépendants de la valeur prise par **le curseur de nombre de clusters**.
- **Le curseur de nombre de cluster :** permet de choisir entre plusieurs clusterings (qui diffèrent donc selon leur nombre de clusters). Il est important de noter que plus le nombre de clusters est élevé plus leur résolution géographique est fine (plus leur étendue géographique est faible). Cependant, lorsqu'un cluster est petit, le nombre de station le composant est faible, il y a donc moins de données disponibles pour ce cluster. Ce paramètre de nombre de cluster peut être vu comme un compromis entre la résolution géographique et le nombre de données disponibles pour l'analyse de chaque cluster.
- **La carte interactive leaflet :** présente les informations choisie dans le **menu déroulant**. Elle est dépendant de la valeur prise par **le curseur du nombre de cluster**. L'utilisateur peut cliquer sur les stations affichées sur la carte. Cela change les informations présentées dans la section **Encadré 4**. Lorsque l'on clique sur une station, toutes les stations appartenant au même cluster sont surlignées en rouge. Cela permet de sélectionner un cluster que l'on pourra étudier plus en détail dans la section **Encadré 4**.

Encadré 4 : Informations complémentaires sur la détection spatiale

Cet encadré regroupe des informations des informations sur une résolution géographique plus fine que dans la section **Encadré 3** ainsi que des informations supplémentaires sur la détection d'anomalie. Il comporte trois éléments :

- **le graphique des concentrations de la stations selectionnée :** lorsque l'on clique sur une station dans la **carte interactive leaflet**, ce graphique donne dans son titre le code identifiant de la station et trace les valeurs de concentrations relevées par la station durant le segment temporel selectionné dans la section **Encadré 1**. Les informations de quantifications de ces relevées présentes sont indiquées.

- **le graphique de front de Pareto des clusters :** dans ce graphique chaque cluster composant le clustering spatial choisi dans la section **Encadré 3** est représenté par un point. Les deux axes correspondent au deux critères qui permettent de juger de l'anormalité d'un cluster. L'axe des abscisses se définit comme l'hétérogénéité des distributions des concentrations des stations composant chaque cluster. Intuitivement, plus l'abscisse d'un cluster est faible, plus les profils de concentrations relevés par les stations de ce cluster se ressemblent. On retrouvera dans les valeurs d'abscisses basses tous les clusters dont les stations n'ont relevé que des données non quantifiées (durant segment temporel selectionné en section **Encadré 1**). Inversement, les clusters comportant des abscisses élevées présenteront beaucoup d'hétérogénéité dans les distributions de concentrations de ses stations. Les concentrations des stations ne se ressembleront pas forcément d'une station à une autre. Pour l'axe des ordonnées, plus un cluster comportera une valeur faible, plus la présence de valeurs élevées et/ou quantifiées dans ce cluster sera faible. Inversement, une valeur élevée sur l'axe des ordonnées indiquera une présence de valeurs élevées et/ou quantifiées. Lorsque l'on clique sur une station dans **la carte interactive leaflet**, le cluster contenant cette station sera surligné en rouge sur ce graphique.

- **les informations complémentaires sur le cluster selectionné :** permettent de résumer des informations sur les données de concentrations comprises dans le segment temporel sélectionné en section **Encadré 1** ET dans le cluster spatial sélectionné dans la section **Encadré 3**. On y trouve les informations suivantes :
  - le nombre de relevés effectuées par les stations du cluster spatial durant la période de temps définie par le segment temporel sélectionné.
  - le pourcentage de quantifications de ces données
  - le nombre de stations composant le cluster
  - le minimum de concentration de ces données
  - la moyenne de concentration de ces données
  - le médiane de concentration de ces données
  - le maximum de concentration de ces données
  - toutes les valeurs de LOQ présentes dans ces données, la valeur de LOQ la plus présente dans les données comporte une astérisque
  - des informations sur la station du cluster ayant le plus grand pourcentage de quantification (son code identifiant, son pourcentage de quantification et le nombre de relevés effectués pendant la période de temps sélectionnée)