Pesticides concentration monitoring from various heteregenous sources of informations

A statistical approach and tool designing

Author: Clément LAROCHE

September 29, 2022

Keywords:
Pesticides, environmental study, spatio-temporal heterogeneity, change points 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

Contents

List of Tables

List of Algorithms

List of Figures

1. Introduction

Studies on environmental data befalls a multidisciplinary domain called environmental science that regroups various fields such as : physics, biology, chemistry, geography, ecology. That important gathering of subjects induced a large collect of data coming from different source of information. As stated in Manly (2008), the emergence of environmental statistics comes from the obvious fact that much of what is learned on the environment is based on numerical data. Three broad types of areas of studies that we believe it is important to state:

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

In recent years, we can cite numerous applications that can be designated as environmental statistics and they span on a very large scope ranging from ice front change monitoring Bunce et al. (2018), emerging marine diseases Harvell et al. (1999), depletion of fossil fuels (Höök & Tang, 2013), vegetation changes (Zheng et al., 2021), temperature evolution (Shi et al., 2022), extreme events occurence monitoring (Zhao & Chu, 2010) and many more areas (Ozgul et al., 2010; Mori et al., 2012).
This thesis proposes a statistical approach that aims at supporting the monitoring activity of phyto-pharmaceutical products. Monitoring the environmental pollution is of great interest for public authorities, important adverse health-effects being well documented nowadays (Khopkar, 2007; Marchant et al., 2018; Nougadère et al., 2014). National health agencies are thus much concerned with monitoring ambient levels and quantifying the concentration of various pollutants in given environmental areas.
Monitoring pollutant in the environment implies to use censors at different locations and that perform samples in different moments in time. Hence, the collected data is spatio-temporal information. Modelling such data is a complex issue, due to several reasons, some intrinsic to the types of data under study, some specific to the data collection process implemented in different countries. Pollutant concentration levels are measured by sensors which have generally detection and quantification limits: the corresponding data are then left-censored. Secondly, the data is usually skewed to the right, with long tails hinting high concentrations. Thirdly, in numerous situations the data is irregularly sampled because of measurement practices, and is often multivariate, since various pollutant levels are monitored. Fourthly, pollution is monitored in various locations, each location possibly using different sensors, yielding a significant spatial heterogeneity.

This thesis focuses on dealing with censored values and spatio-temporal heterogeneity. This demands a procedure that articulates different models and methods. The general princple is to find time periods and spatial areas where the informations are the most homogeneous possible using the most coherent datasets possible. Once these moments and zones are identified, we aim at detecting the most anomalous zones in these time periods. The manuscrit is organized as follows:

- **Chapter 2** is an introduction to the French national agency in charge of monitoring pollution data. A short description of the agency and its missions are given with an extensive description of the data available. Several datasets from different sources of information are useful to the analysis of environmental pollution. A first exemple of how to combine these informations to extract some information in the last section.

- **Chapter 3** makes an inventory of methods that are useful in the analysis of environmental pollution. The analysis of such data is difficult in presence of high spatio-temporal heterogeneity. We are looking to cut the time series into more homogeneous subdivisions. This objective befalls the change-point detection field in the litterature. We provide review in such methods in this Chapter.

- **Chapter 4** builds a specific parametric change-point detection method. We are looking for an adapted method for the problems we are facing. We study the effect of censorship to show that it does not prevent to find change-points in a signal. We also derive an optimization procedure that is suited to some modeling configuration. Simulation experiments are led to compare it with a state of the art non parametric method that is also adapted to censored data.

- **Chapter 5** provide the spatial analysis of concentration data in the environment. It uses the results of the change-point detection method developped in Chapter 5 on the temporal dimension. Using spatial clustering and anomaly detection methods, we manage to extract some informations useful to assist experts in the environmental pollution monitoring mission.

- **Chapter 6** describes the elaboration of an interactive application that displays the results of our procedure. This application serves an operationnal purpose. It is specifically designed for the experts working in that area of expertise

2. Application context

Contents

One of the most recent political answer brought to tackle environmental issues in Europe took form under the European Partnership for the Assesment of Risks from Chemicals (PARC) *Web ressource: European commission* (n.d.). This partnership involves 28 different countries. This new project received a favourable assessment by the European Commission in January 2022 and has started on the 1st of May 2022. The main objectives of PARC are to promote European cooperation, to advance research, to increase knowledge about the risk assessment of chemicals and to train the corresponding methodological skills. Close cooperation between authorities and research will facilitate the translation of research results into regulatory practice.
The French Agency for Food, Environmental and Occupational Health and Safety (ANSES) is not only the main French actor in this partnership, but also the coordinator of the entire partnership. This work was funded by the ANSES and aims to support the agency in its mission on French territory. This chapter describes the overall context and the specific problems arising from the characteristics of the data used. The chapter is organized as follows: In 2.1 we introduce ANSES, in 2.2 we detail the task of pesticide monitoring, in 2.3 we focus on the characteristics of pesticide measures, in 2.4 we describe other sources of information of interest, and in 2.5 we define the objectives sought.

2.1. ANSES presentation

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

2.1.1. Missions

- **Research activities:** the agency contributes to the progress of new scientific knowledge on the exposure of humans, animals, plants, and the environment to various hazards and risks and is tasked with improving their surveillance. 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, whether in the natural environment or in the production chain. This mission also regroups the activities health monitoring and alert. The agency takes part in epidemiological surveillance platforms on animal health, plants health and food chain safety. This mission also tasks ANSES to coordinate five different monitoring systems that covers the following areas: toxicovigilance, surveillance on food supplements, phytopharmacovigilance which will be further investigated in 2.2, vetenary pharmacovigilance and surveillance and prevention of professionnal pathologies.

- **Risk assessment:** ANSES responds to society's questions about potential risks arising from the consumption of food, the use of certain products or technologies, professionnal activities, or pollution of various environmental compartments (e.g., air, water, or soil). Given the complex risks, the agency has developed a working methodology that brings

together many disciplines to provide the most complete response possible. The risk and efficacy assessment of veterinary medicines and phytopharmaceutical products for human and animal health and the environment also falls within the agency's remit. The goal is to define management measures to handle such risks. In particular, ANSES is responsible for granting marketing authorizations for such products and thus also has the authority to withdraw products at the national level *Web ressource: Anses decision site* (n.d.).

- **Public and environmental protection:** 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 public access to reliable, independent, and multidisciplinary scientific information. The agency's role is to respond flexibly to already known or emerging, short- or long-term sanitary risks. In other words, the goal is to identify and make recommendations on all emerging signals as quickly as possible, even in times of crisis with scientific uncertainty. The task is then to reduce the degree of uncertainty when possible. Recommendations are based on all available knowledge, whether it was generated by the agency itself or by its partners.

2.1.2. Means of action

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

- learning the characteristics of pathogens (such as fungi, bacteria, viruses or parasites), macroorganisms (such as insect pests or invasive plants) and chemical contaminants.
- detect them using state-of-the-art analytical methods.
- monitor them using powerful epidemiological methods.
- understand the impact of animal husbandry on animal welfare and health.
- 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 ressource: Anses laboratories mandates* (n.d.).
In addition to its research activities, the ANSES is also at the center of a network of partners. Given the scope of the areas it covers, the agency does not have the resources to collect data on all the topics it is asked to address. Therefore, for each topic, it conducts discussions with

other organizations that are likely to provide interesting sources of information. For example, each monitoring system coordinated by the agency mobilizes a different set of partners. We will see the different datasets that those partners bring in sections 2.3 ans 2.4.

2.1.3. Specific organisation

Many national agencies are counterparts of ANSES. They all participate in the PARC partnership. Each shares data collected at the national level, enabling research studies at the European level. However, each agency is dependent on the network that collected the data it stores and on the internal policies of the country to which it belongs. This leads to heterogeneity on different topics and at different levels.
For example, internal policies influence the list of monitored products, and some substances are not studied in certain countries because they are not even approved for sale. This leads to a patchwork of different susbtances lists at the European level as shown in Baran et al. (2022). Some heterogeneity is also observed at the national level. For example, in France, it was determined that drinking water quality data should be collected at the regional level 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 2.3.
Thus, the structure of the monitoring system is country-specific. This is especially true for the pesticides monitoring system, on which we will focus on 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 characterized by the following questions:

- what was the degree or intensity of exposure?
- how long and how regularly does the exposure occur?
- in what manner does the exposure occur? (Is it skin contact, ingestion?).

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 diabete or asthma). The outcome depends on the characteristics of the exposure and the characteristics (like age or immunity) of the animal, human or plant population under study.
ANSES is tasked with studying and monitoring health risks caused by various factors. In particular, the agency is charged with monitoring health risks associated with chemical agents. The pesticide surveillance mission can be formally defined as a surveillance system that collects and evaluates monitoring 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 soon as possible in order to protect the health of living organisms and ecosystems. Pesticide health hazards are well referenced by the Agency and available in the AGRITOX database. Different types of exposures can be distinguished depending on the population affected by the monitoring. We give two different specific cases of exposures with different populations. The first case is about monitoring the health risk of professional farmers. They are regularly exposed to pesticides that they apply. The exposure is then long-term and the likely routes by which they could come into contact with the substance would be inhalation or dermal contact. The second case involves monitoring the health risk to aquatic fauna. Exposure to pesticides may result from the effects of water runoff following pesticide application, which could lead to dispersal into the river system in the area. Since 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, depends on external factors, independent of the intrinsic properties of the substance. One example would be meteorological conditions during the study period. Another example would be to consider the environmental characteristics that are likely to influence the diffusion of the substance. For example, the diffusion of a chemical product in a river system might be influenced by the riverbed composition or the width of the river. Identifying regions where those characteristics are fixed would be idea for studying the exposure, the hydro-ecoregions (HER) provide a good took to work with in the river system example.
Anses does not itself collect the data it needs to fulfil its phytoharmacovigilance mission. It relies on its partner network to obtain data sets of interest. The decision to add a new source of information first requires a discussion of the coherence of the use of that source of information to provide an answer to the problem under study. 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.

This results in collecting spatio-temporal informations. The data sets that provide information on pesticide concentrations and use are presented in detail in section 2.3.

2.3. Pesticides measurments 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 of a substance.

2.3.1. Direct measurments characteristics

In this part, the characteristics of pesticides direct measures from sampling stations are studied. The following characteristics are not found systematically, but are very common when it comes

to concentration measurements.

Chemical precision limits in measurments

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

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

These limits are determined by the sensors that the sampling station is equiped with. It happens that geographical areas are covered by stations that do not have the same equipment. In this case, there are several LOQ values within the samples taken in that area. In the case of surface waters, for example, the contracts for the selection of monitoring laboratories are awarded by the water agencies. These agencies, six in number, cover an area larger than the French administrative regions. This means that if the scale of an administrative region is taken as the basis for a study, this region may fall under the jurisdiction of two different water agencies and therefore the measuring instruments may be different. Moreover, the same station may change its measuring equipment over time. Station equipment contracts are renewed periodically, but renewal does not guarantee that the same equipment will be maintained. All those caracteristics are illustrated in Figure 2.1. These two limits of accuracy mean that concentration data are left-censored. In Chapter 4, we will show which method was chosen to handle this type of data in this thesis. Several methods have been developed to handle this type of data. We can cite the imputation of values to replace the LOQ values present in the set of concentration values, the use of the maximum likelihood estimator or the Kaplan-Meier estimator (see Gillaizeau et al. (2020); Croghan & Egeghy (2003)).

Irregular sampling

The second feature is a direct consequence of Section 2.1. We mentioned that each country has organised its own monitoring system. Some features are country specific and have an impact on the raw data of the collected samples. Sampling in France is not necessarily done at regular intervals, as can be seen in Figure 2.1. No sample was made during the year 2019. This is also true for water monitoring data. This is a peculiarity 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 regular 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, and a sample is taken every six hours and automatically stored on a central server after analysis. We would also like to point out that strategies other than regular sampling, such as grab sampling Novic et al. (2017) could explain the irregular sampling rythm in French surface water quality. Grab sampling consists of obtaining as accurate a picture as possible of surface water quality in a short period of time

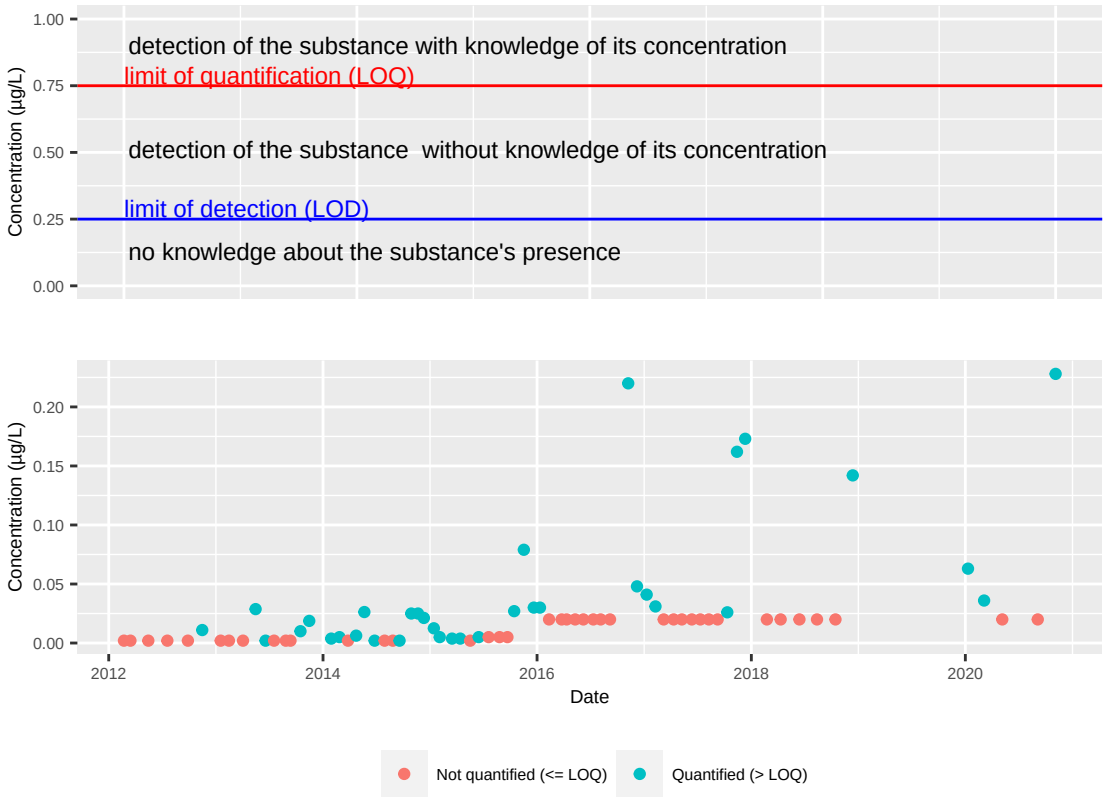

Figure 2.1: Censorship illustration. The top Figure sums up the limits of measurements effects. The bottom Figure shows the consequences of censorship on the samples of a station located in the Centre-Val de Loire region. This station changed its equipment in 2016, the LOQ change values.

(and in a limited area). If the operations of grab samples and periodic monitoring operations of a station are reported in the same database, irregular sampling may also occur.

Spatio-temporal heterogeneity

Another main chgaracteristics mentionned in Baran et al. (2022) is the spatio-temporal heterogeneity in the data-sets. Figure 2.2 shows the temporal heterogeneity. For two spatially neighbouring stations the measured values are not synchronous in time. It can also be seen that the stations take very few measurements and do not take the same number of measurements. In Figure 2.2 we see that it is not possible to compare the measurements of the two existing stations. The stations made samples in 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.

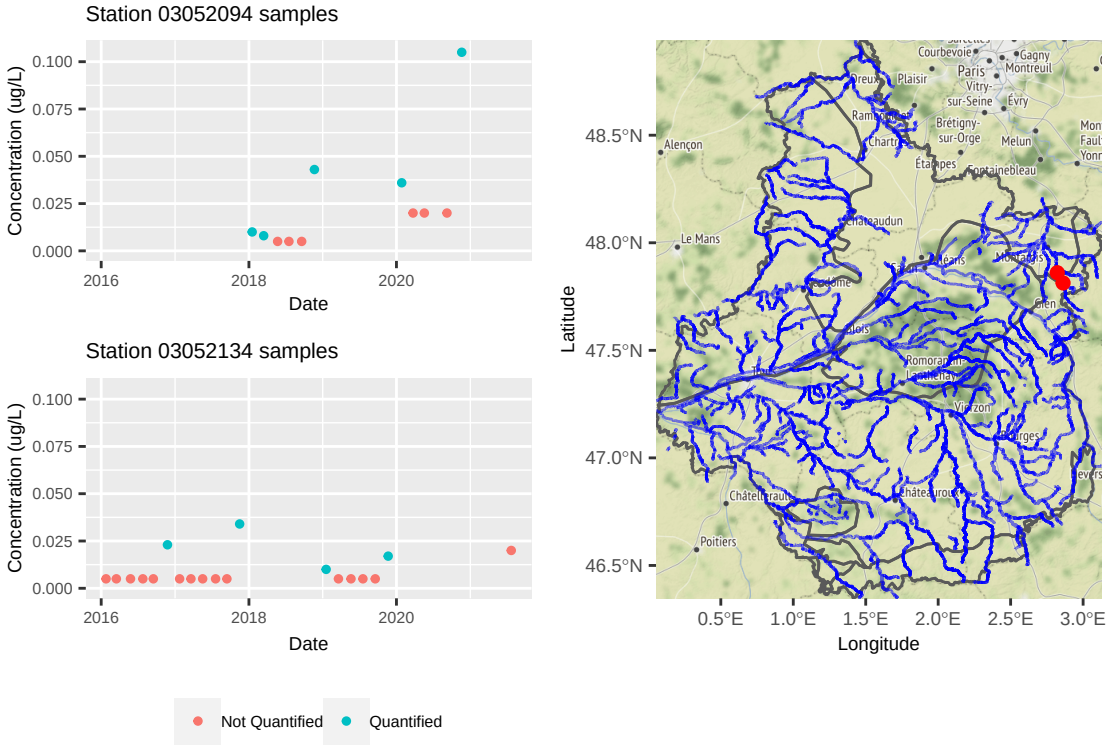

Figure 2.2: Spatial and temporal heterogeneity in sampling. The Figures on the left represent all the samples of two neighbouring stations. The map on the right shows the position of those stations.

Figure 2.3 illustrates that alongside temporal heterogeneity induced by the stations different sampling rhythms, spatio temporal data are not distributed in a homogeneous way over the territory. The concentrations value seem to completely differ according to their spatial area of origin. Thus, the data are heterogeneous in space. Figure 2.3 also illustrates that the

distribution of concentration values can drastically change over time. Looking at the samples of station 03189000, there is a break point just before the year 2015 in the station concentration values. The same can be said for the year 2016 in Figure 2.1. Thus the data are heterogeneous in space and in time.

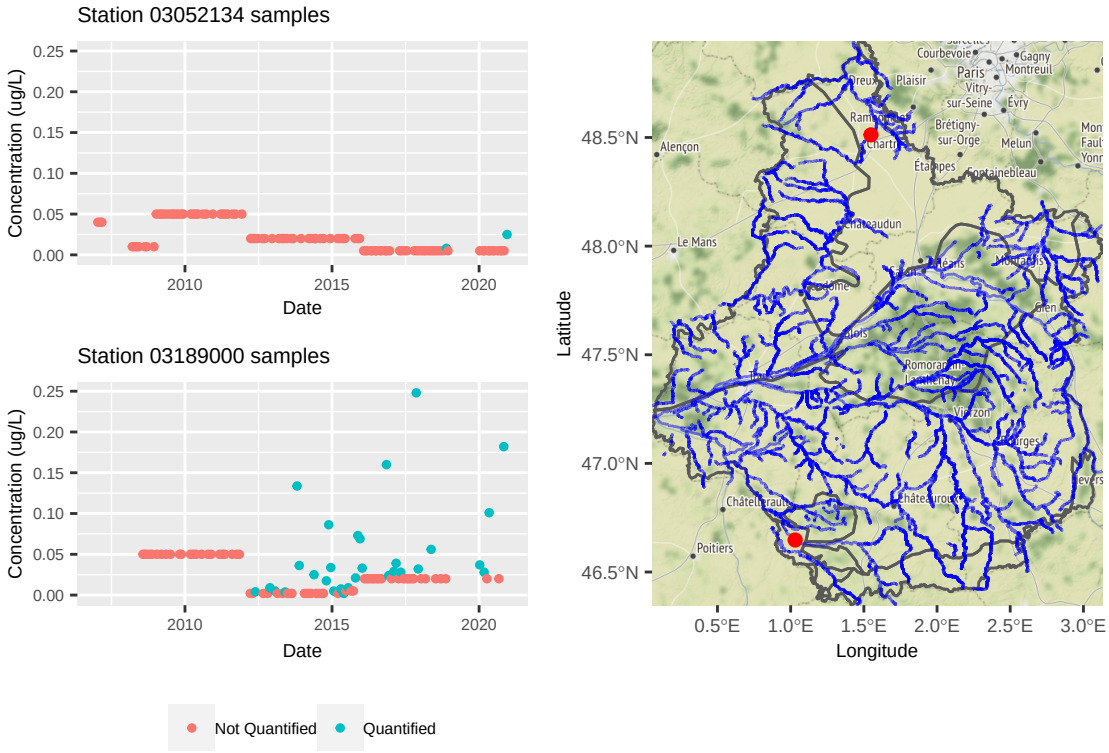

Figure 2.3: Spatial and temporal heterogeneity in distribution. The Figures on the left represent all the samples of two stations. The map on the right shows the position of those stations.

2.3.2. Indirect measurments information

Surveys on farming pratices

Additional clues to the substance presence can be obtained from surveys of agricultural practices. They are conducted in the form of a questionnaire and are used to describe and characterize how farmers operate on their land. These studies are very specific and focus only on certain types of crops. Three topics are addressed in the questionnaires. The first captures general information about the farm, such as commitment to pesticide use reduction or agroecology. The second questionnaire is designed to capture all technical operations on the farm plot. In other words, we examine the structure of the plantation, its preceding crops or its irrigation. Finally, the use of pesticides on the whole farm is studied. The questionnaire goes over criterias

such as the type and settings of the sprayer for the substance or the handling and protection of the user. Figure 2.4 shows what kind of questions to expect in this survey[1]. In summary, farming practice surveys provide much qualitative (rather than quantitative) information about the source of substance emissions. However, these surveys are conducted on an ad hoc basis and cover only certain types of crops. Only the results of the statistical departments of Agricultural ministry analyses *Web ressource: Agreste source* (n.d.) are available however the raw data is not.

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 2.4: Question extracted from the 2019 survey destined to vitivulture. The question asked is about the fertilization and protection of the crops and the use of any phyto-sanitary product.

Substances sales databank

The use of a substance can also be indirectly seen in the sales data of crop protection products. The National Bank for the Sale of Pesticides by Authorized Distributors (NBSD) Office français de la biodiversité and Système d'Information sur l'Eau (2021) lists and archives all such data. For the same reasons of anonymity, geographically fine resolution information is not available. The most accurate resolution corresponds to postal codes. It is the same for the temporal resolution that is not finer than the yearly resolution. Rhis data set does not indicate the location and date of use of the substance. A purchaser may well be in a different location than the place of use of the substance they just purchased. Nevertheless, sales give a general indication of the intensity of use of a substance. A sudden increase in sales of a product may mean that its use is increasing in that area.

Crops cartography

Specific pest species can be observed for each crop type. Mapping the crop types in an area can therefore provide a preliminary idea of the areas and periods of application of the substance being monitored. Some of this information is available in the graphical land register (GLR) *Web ressource: IGN data.* (n.d.). This database corresponds to the application forms used by farmers to obtain financial aid under the Common Agricultural Policy of the European Union

[1]Document in French, full document in *Web ressource: 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.

(CAP). To be eligible for these grants, the crops grown on the plots must be declared. This dataset is a partial information, since asking for CAP funds is not mandatory. Therefore, the owners of the crops who have not applied for aid are not present in the database. Moreover, this register is renewed every year. It is possible that the information for certain parcels is not included in all annual editions of the GLR. Crops maps of barley and wheat are diplayed in Figure B.3 of Annex B.3.1.

Adverse effects databases

The last example of data that can highlight clues of a substance usage are the databases for monitoring potential adverse events. They consist in medical registries providing informations on human and animal health. For human health, several information sources 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 he/she has health problems. This organization collects data through spontaneous reports from agricultural actors or during scheduled visits by nurses or doctors.

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

- the poison control centers are involved in adverse effect surveillance. They provide toxicovigilance information on toxicovigilance for the entire population. Much information about acute health problems comes up through these information channels.

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

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

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

2.4. Example of additional useful data

We discussed in Section 2.2 the exposure factor in monitoring a health risk. 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. Since an exhaustive list of all possible data sets would prove lengthy, this section provides examples of interesting additional data sets for surface water and air quality monitoring.

2.4.1. Surface water quality

In surface water quality monitoring, stations are positionned on streams of water or lakes. Their precise GPS location can bring insight on how a substance could diffuse once introduced into the surface water system. This information is made available by the National Institute of Geographic and Forest Information (IGN) services in the BDTOPO database. Figure 2.5 displays the river system of the Centre-Val de Loire French region with the positions of all stations. With this information at hand, stations concentrations can be directly compared according to their distance in the river system.

However, we mentionned in Section 2.3 that an individual station does not provide much measurments. Therefore, to derive information from these data, one can work on a different resolution than that of the station. Thus, there is a trade-off between the number of data available to make a statistical statement about a spatial area and the spatial resolution accuracy. Another interesting level of resolution that was briefly mentionned in Section 2.2 are the hydroecoregions (HER). They are geographic units in which hydrographic ecosystems share common characteristics. The criteria by which they are delineated combine characteristics of geology, terrain, and climate Wasson et al. (2002). INRAE services provides such information. Pooling all samples from the HERs can provide a satisfactory level of aggregation but it would be at the expense of the spatial resolution. Figure 2.5 shows how the stations are distributed in the HER.

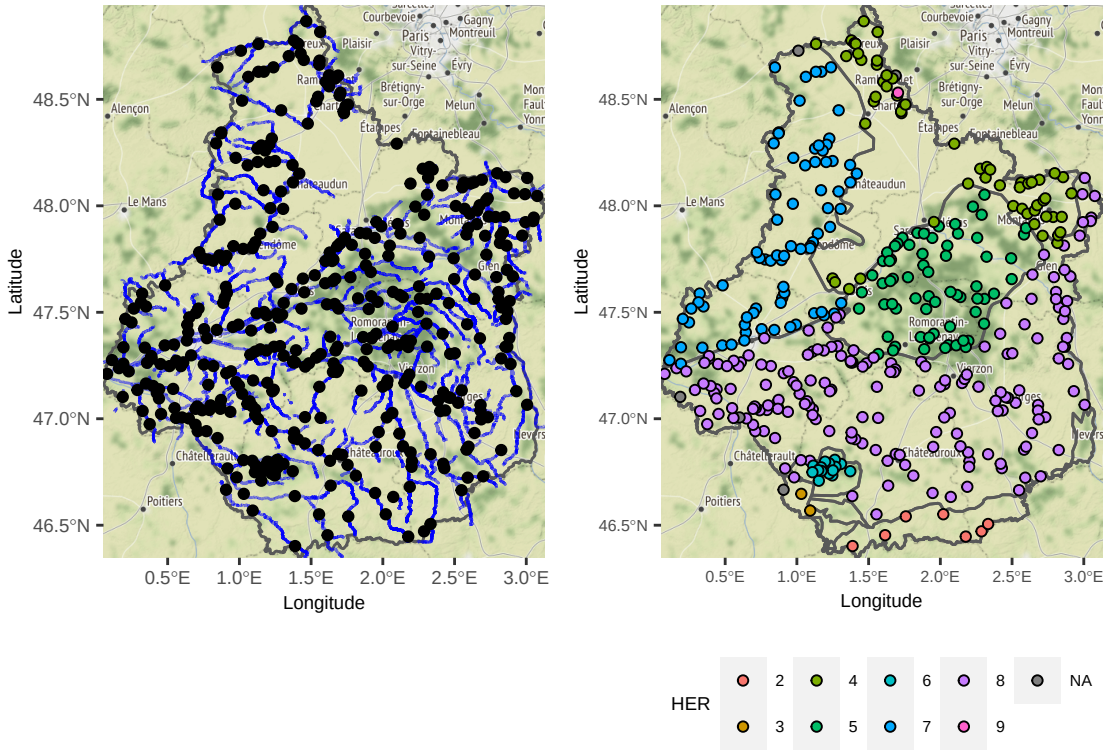

Figure 2.5: Stations monitoring water suface in the Centre-Val de Loire french region. Two different geographical resolutions are represented. The underlying hydrographic network linking all stations is plotted on the left, the stations are colored according to their hydroecoregion on the right.

2.4.2. Air quality

Meteorological data is an important data set for monitoring air quality, especially any information that can be found about wind (wind direction, wind strength, etc.). Historical weather records are now available as open data on the Météo France website *Web ressource: Météo-France data (SYNOP).* (n.d.). Figure 2.6 illustrates the cross-referencing of data from air quality monitoring stations with meteorological data. Note that in this example we chose to represent stations within the study region, but information from outlying areas can also provide interesting informations on air pollution in the selected area. This example also shows that concentration monitoring tasks depend on the application context. The coherence of any data set included in the analysis must be discussed.

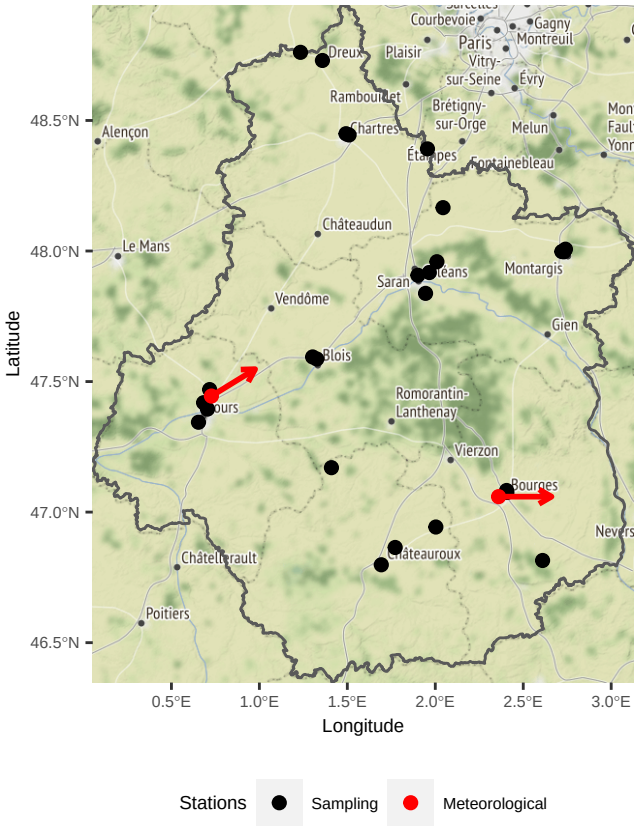

Figure 2.6: All active stations measuring air quality on March the 1st of 2021 coupled with meteorological stations active that day. The main wind direction and speed measured that day is mapped with the red arrows.

2.5. Surveillance of pesticides data

This section proposes an initial exploratory approach to monitoring pesticide concentrations using data from Sections 2.3 and 2.4. The goals set in 2.2 may be difficult to achieve because of the specific French administrative organization discussed in 2.1. The spatiotemporal nature of concentration data requires appropriate techniques for their representation, which were reviewed in Andrienko et al. (2003); Cressie & Wikle (2015); Maimon & Rokach (2010). 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 domain. experts. This includes the development of visualization tools. Several methods are available for visualizing spatiotemporal data. This section presents some visualization techniques and discusses their limitations.

The spatial map plots or iterative maps Andrienko et al. (2003) are an effective way to extract information from these data. They consist of maps of the same phenomenon at different times, as in Figure 2.7. There is a clear seasonal pattern in this figure. We can conclude that prosulfocarb is applied in autumn, and this information is confirmed 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 account for years in which 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 summary indicator of quantification rate is used. Nothing is known about the maximum or average concentrations.
Another conventional representation is the display of information with a map animation Andrienko et al. (2003). In this method, the information displayed on the computer screen is updated depending on the selected spatial area. Figure 2.8 shows a practical example of this technique. Concentrations of prosulfocarbe in the Centre-Val de Loire French region of France are displayed according to the selected HER region. Two limitations arise with this visualization. The first 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 spatial resolution in Chapter 5. This presentation also raises the question of the choice of temporal resolution. All samples available in the study period are shown in Figure 2.8. There is a clear break in the three series around 2015. There appears to be a change in concentration regimes. A finer temporal resolution combined with a comparison based on statistical inference of the selected geographic regions could better help experts in the interpretation. Chapters 3 and 4 will address the issue of temporal resolution.

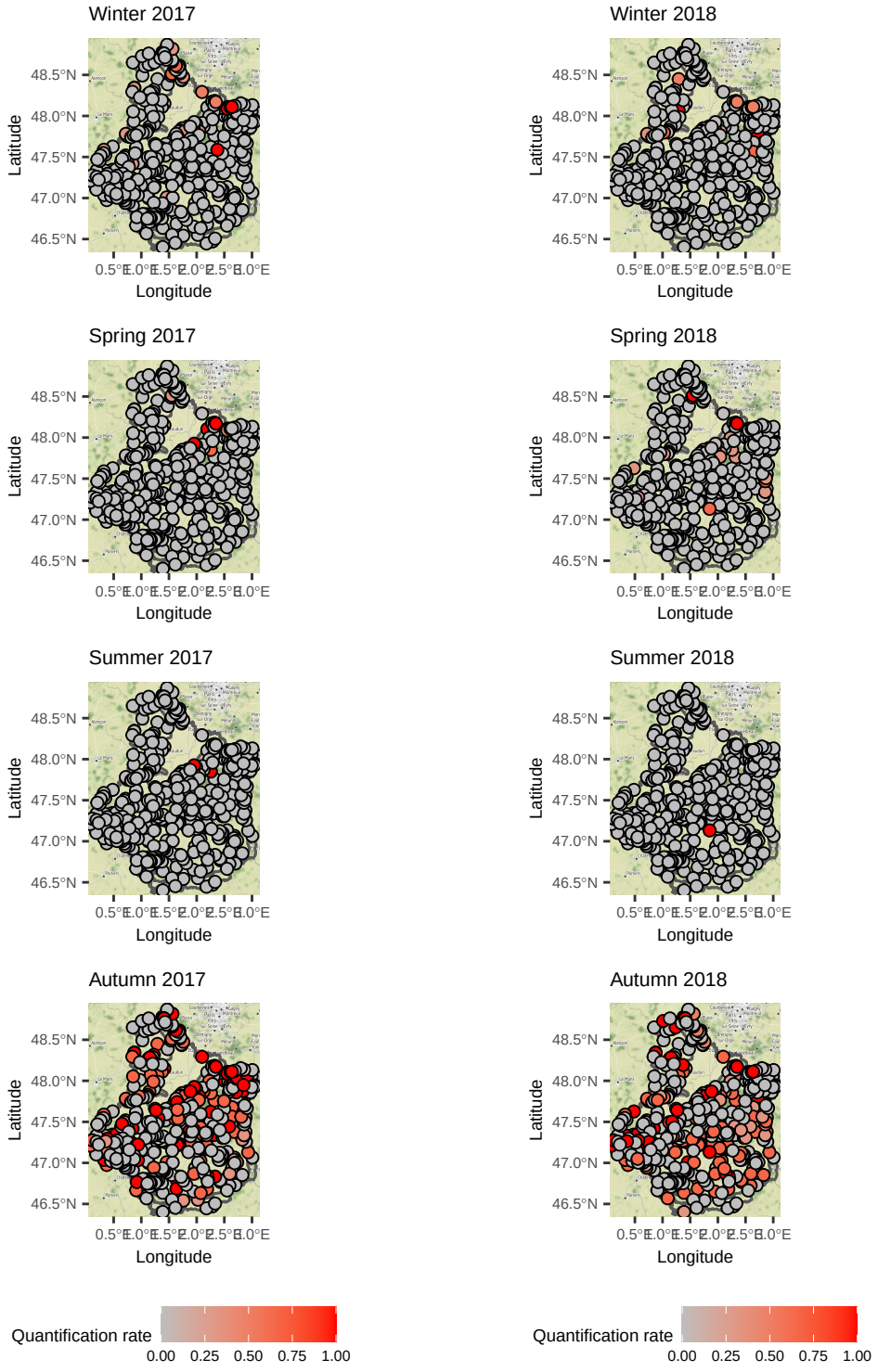

Figure 2.7: Spatial maps in time. Prosulfocarbe's quantification rate of each station was computed for each season of 2017 and 2018.

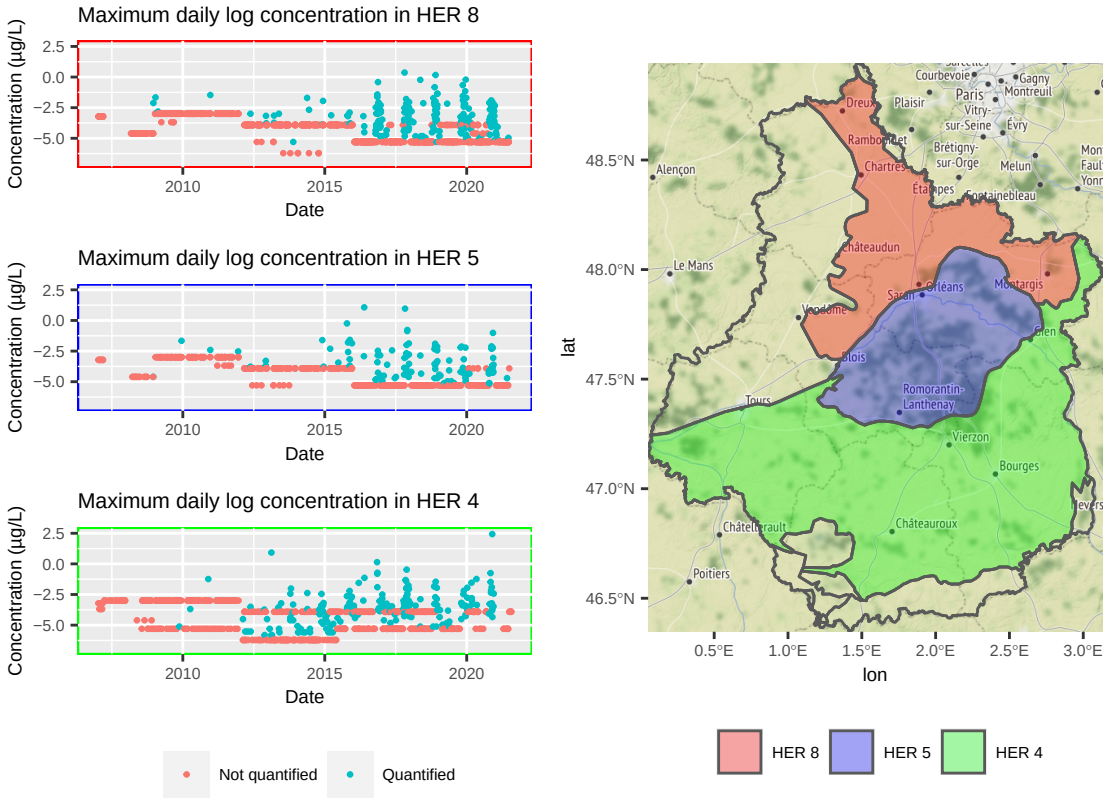

Figure 2.8: Time series plot of HER 8,5 and 4 daily maximum concentrations. The log scale was used for a easier visualization.

3. Researching homogeneous temporal periods in time series

Contents

Although the recommendations for the use of the substances are formulated in the marketing authorization issued by ANSES *Web ressource: ephy catalogue* (n.d.), their practical use is not subject to control by the agency. Cultivation practises depend on meteorological conditions and professional habits. This partly explains the spatiotemporal heterogeneity. Our goal is to find homogeneous time periods and geographic areas and make a statistical comparison of these regions in these stable time periods to support the expert analysis. In this chapter, we focus on identifying stable temporal patterns using change point detection methods. This mathematical area is covered by several surveys Truong et al. (2020); Basseville & Nikiforov (1993); Bardet, Jean-Marc et al. (2020). We have focused on methods that seem appropriate for the application domain of phytopharmacovigilance. For example, this work will focus only on the offline methods that we will develop in the following sections. This choice was motivated by the speed of data collection and storage of pesticide concentrations data. Nevertheless, for readers who wish to refer to it, we can state that there is no shortage of online detection methods in the literature S. Liu et al. (2017); Y. Li et al. (2021); Höhle (2010); Ranganathan (2010); S. Li et al. (2015).

3.1. Model and cost functions

We describe the most general configuration of a change-point model appropriate for concentration data. 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. Thus, the indices in Y_i are only indicators of the order of occurrence in the sample and not of the observation times. Some properties (trend, mean, variance, etc...) of the signal $\boldsymbol{y}$ are supposed to change at the K^* times points $\tau_1^* < ... < \tau_k^* < ... < \tau_{K^*}^*$. We use the following convention, let $\tau_0^* = 0$ 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. The goal is to identify the data segments in which these properties are stable. We denote $y_{u:v}$ as a segment of the signal from the u-th coordinate to the v-th.
According to the nomenclature proposed by Truong et al. (2020), change point detection methods operate on a cost function W. This function associates a cost to the segment it is evaluated on. Intuitively, the more properties (on which changes are investigated) of the segment $y_{u:v}$ are homogeneous, the lower the cost $W(y_{u:v})$ is. We define, for any $y_{u:v}$, $\mathcal{T} = \{\tau_1, \dots, \tau_K\} \subset \{u, \dots, v\}$ a set of ordered indices and $|\mathcal{T}|$ its cardinal. Implicitly, we define $\tau_0 = u$ and $\tau_{K+1} = v$. Although this notation is often used for the full signal $\boldsymbol{y}$, we will always indicate the data segment from which $\mathcal{T}$ is drawn in the parameters of the functions using the notation $\mathcal{T}$. The total cost $\mathcal{C}(\boldsymbol{y}, \mathcal{T})$ associated with a segmentation defined by $\mathcal{T}$ 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 the knowledge of the number of change points K^* that occurred in $\boldsymbol{y}$, the change point problem can be posed as an 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 cost function 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 on which they are based. We give a non-exhaustive list of cost functions for each inference.

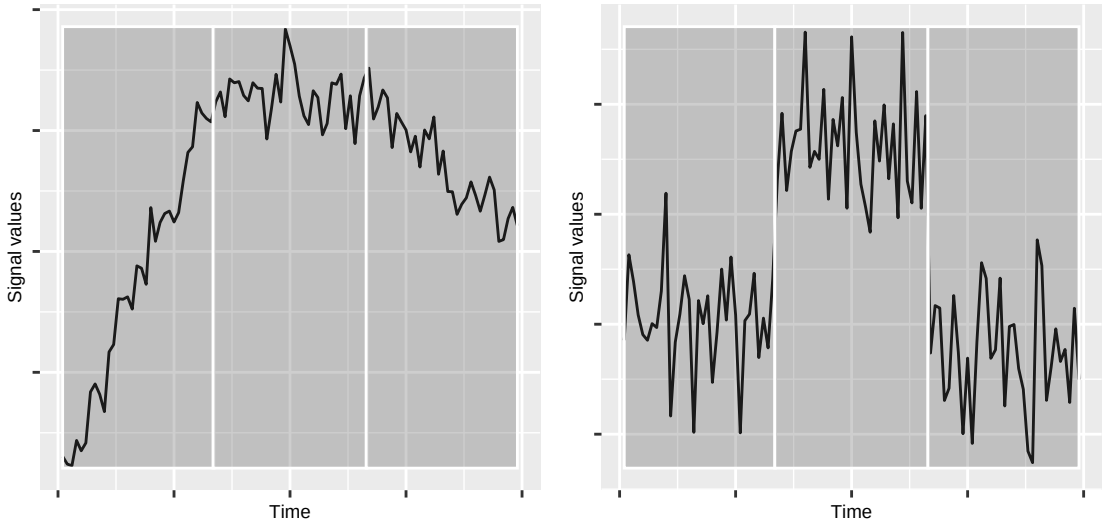

Figure 3.1: Examples of types of change point detection. The figure on the left illustrates changes in trend whereas the figure on the right illustrates changes in mean.

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).
A first classical cost function is based on the maximum likelihood estimation. In this setting, the observations located in the k-th segment is supposed to be following a distribution Q depending on a vector of parameters θ_k^* with $\theta_k \in \Theta$ being a compact subset of $\mathbb{R}^p$. More formally, we have that:

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

with f being the density function of distribution Q. In other words, we suppose that all observations emanate from the same distribution Q but the values of θ_k^* change abruptly at each change-point τ_k^*. The cost function used to evaluate segments in this context is the negative

log-likelihood. Hence, for a segment $y_{u:v}$ with $u < v$, we can write:

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

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 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 simpliest 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 is supposed can be written as:

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

where $\theta_k^* \in \mathbb{R}^p$ are the regression parameters and ϵ_t is the noise of the signal. The adapted cost function in this configuration uses the least squares estimation and is expressed as:

$$W(y_{u:v}) = \min_{\theta\in\mathbb{R}^p} \sum_{t=u+1}^{v} (y_t - x_t'\theta)^2$$

This modeling is perfectly suited for the detection of the changes in the left example of Figure 3.1.

3.1.2. Non-parametric inference

The cost function for a segment can also be adapted for nonparametric statistical inference. Several strategies have been developed in the literature over time. These include the nonparametric maximum likelihood method 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 $R_i = \sum_{j=1}^n \mathbb{1}(X_j < X_i)$. Moreover, we note $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i < t)$ the empirical cumulative distribution function (c.d.f.). The cost function is derived from the Wilcoxon/Mann-Whitney rank criterion. This is equivalent to running the test under the following assumptions:

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

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

$$U_n(t) = \frac{2}{\sqrt{nt(n-t)}} \sum_{i=1}^{t} \left(\frac{n+1}{2} - R_i\right) \tag{3.3}$$

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

$$S_n(t) = \hat{\Sigma}_n^{-1} U_n^2(t), \tag{3.4}$$

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

$$W(y_{u:v}) = -(v-u)\hat{\Sigma}_n^{-1}\overline{R}_{u:v}^2, \tag{3.5}$$

where $\overline{R}_{u:v} = \frac{1}{v-u}\sum_{i=u}^{v} R_i$ is the average rank of $y_{u:v}$. This methods allows the identification of segments where the ranks of the obsevations are homogenous. 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 likehood method finds similar results to 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. Estimating an unknown number of change points

Various search methods for finding breakpoints have been described in the litterature. They can be distinguished according to whether they provide an optimal solution to the problem of the research 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 optimal methods. This choice is motivated by the size of the datasets we apply change-point detection on. The number of samples is still in a reasonnable range to obtain satisfying computationnal times.
Since the pesticides uses are supposed to be spread at regular times during the years, we could expect a seasonnal 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 to be certain of this assumption. This section discusses the case where the number of change points K^* is unknown. We present different ways to estimate the number of breaks and their positions.

3.2.1. Optimal partitioning method

The optimal segmentation algorithm brings an answer to problem 3.2. This is the "brute-force" search method. It necessits to compute the cost of all possible segments $y_{u:v}$ with $u < v$. With a fixed K number of change-points, one can recursivly solve the optimization problem. The recursion comes from the following relationship:

$$\min_{|\mathcal{T}|=K} \mathcal{C}(\boldsymbol{y}, \mathcal{T}) = \min_{t \leq n-K} \{W(y_{1:t}) + \min_{|\mathcal{T}|=K-1} \mathcal{C}(y_{t:n}, \mathcal{T})\} \tag{3.6}$$

In other words, given all possible segmentations of all sub-signals $y_{t:n}$ in $K-1$ segments, one can compute the optimal segmentation of the whole signal $\boldsymbol{y}$ in K segments. This results in a computationnal cost in order of $\mathcal{O}(Kn^2)$ Haynes et al. (2017). Algorithm 1 shows the implementation of 3.6.

Algorithm 1 Optimal partition algorithm:

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

A practical aspect of this implementation is that running Algorithm 1 for a given K_{max} provide the results of all optimal segmentations for all $K \leq K_{max}$. The downsides reside in its computationnal cost which is expensive and 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 partitionning. We give two examples of how to proceed.

- **Penalizing the cost:** as mentionned 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 addtionnal cost one must pay each time a break is decided in the signal $\boldsymbol{y}$. This gives the new optimization problem:

$$\min_{\mathcal{T}} \{\mathcal{C}(\boldsymbol{y}, \mathcal{T}) + pen(\mathcal{T})\} \tag{3.7}$$

  Once the optimal partitionning method has been applied with maximum number of change points K_{max}, we obtain the resulting segmentations $\{\widehat{\mathcal{T}}_1, \ldots, \widehat{\mathcal{T}}_{K_{max}}\}$ and their associated

costs $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_1), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}})\}$. We can apply the penalization procedure on these costs and estimate K by selecting the minimal penalized cost:

$$\widehat{K} = \arg \min_{K \in \{1,\ldots,K_{max}\}} \{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_1) + pen(\widehat{\mathcal{T}}_1), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}}) + pen(\widehat{\mathcal{T}}_{K_{max}})\} \tag{3.8}$$

$\widehat{K}$ 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). We discuss our choice in Section 3.3.

- **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 bipartite linear model on the costs $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_1), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}})\}$. 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 $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_1), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_K)\}$ and $\{\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_K), \ldots, \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}})\}$. A `R` based algorithm is provided in Algorithm 2.

Algorithm 2 Elbow method algorithm

input : the segmentations cost resulting from optimal partitionning $\mathcal{C}(\boldsymbol{y}, \mathcal{T}_K)$ for $K \in \{1, ..., K_{max}\}$

initialisations : Initialize $C \leftarrow (\mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_1), ..., \mathcal{C}(\boldsymbol{y}, \widehat{\mathcal{T}}_{K_{max}}))$,
Initialize $slope \leftarrow (0, ..., 0)$ a $K_{max} - 2$ length vector.
for $k = 2, ..., K_{max} - 1$ **do**
  $ml1 \leftarrow \texttt{lm}(C(1:k) \sim 1:k)$
  $ml2 \leftarrow \texttt{lm}(C(k:K_{max}) \sim k:K_{max})$
  $slope(k-1) \leftarrow \sum ml1\$residuals + \sum ml2\$residuals$
end for
$CP \leftarrow (2 : K_{max} - 1)(which.min(slope))$
output : the optimal number of changes CP.

Appplication of optimal partitionning methods can be found in Rigaill (2015); Lavielle (1999); Perron et al. (2006)

3.2.2. PELT algorithm

Problem 3.7 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 parameter β takes positive values. It corresponds to the cost assigned to a breakpoint. Intuitively, the higher the penalty value is, the lower the number of change points detected is. Using PELT, one can sequentially go through the signal $\boldsymbol{y} = \{y_s\}_{s=1}^{n}$ and obtain a set of potential breakpoints $\{\tau_0, ..., \tau_m\}$ for each index $s \in 1, ..., n$. Then, one should proceed to eliminate candidates from this set using a pruning rule involving the penalty value β. 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

$$\min_{\mathcal{T}} \Big[\mathcal{C}(y_{1:t}, \mathcal{T}) + pen(\mathcal{T})\Big] + W(y_{t+1:s}) \geq \min_{\mathcal{T}} \Big[\mathcal{C}(y_{1:s}, \mathcal{T}) + pen(\mathcal{T})\Big], \tag{3.9}$$

holds, then t can never be the last changepoint prior to n. We introduce some additionnal notations to simplify the algorithm writing:

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

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 $R_{\tilde{t}+1} = \big\{t \in R_{\tilde{t}} \cup \{\tilde{t}\} | F(t) + W(y_{(t+1):\tilde{t}}) + \beta \leq F(\tilde{t})\big\}$
end for
output : the vector of change-points CP.

Using the notations, the role of parameter n_{min} writes as $\min_{i \in \{0,...,K\}} |\tau_{i+1} - \tau_i| \geq n_{min}$. It has a direct influence on the segmentation and acts as a compromise between the segmentation resolution and the cost function precision. Small values of n_{min} lead to the detection of a large number of change-points. However, the statistical validity of these changes can be questionned. Large values of n_{min} ensure satisfying convergence of the cost function, but one risks to miss changes that occured at near 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 constitues a major improvement compared to the optimal partionning method. However, as mentionned 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 $\boldsymbol{y}$. Diverse strategies to calibrate β exist as the BIC criterion which is widely used Yao (1988); Faure et al. (2016); Shi et al. (2022) or some data-driven heuristics Birgé &

Massart (2006); Baudry et al. (2011); Bardet et al. (2012); Arlot & Massart (2009). Despite all that, there is no quantification possible for this parameter: we cannot predict the number of breakpoints resulting from a given fixed value of β. Thus, we don't know if the change point model resulting from the choice of β is over or under fitting the signal $\boldsymbol{y}$. We need a more exploratory approach to tackle this problem.

3.3. Exploratory research of segmentations

We are looking for a compromise between the exhaustivity of the "brute-force" optimal partionning method and the computationnal cost of PELT. The algorithm CROPS: Changepoints for a Range Of PenaltieS algorithm Haynes et al. (2017) allows to search a range of penalties $[\beta_{min}, \beta_{max}]$ and to find penalty values wihtin that range associated with new segmentations. The proccess 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 changepoints of $\boldsymbol{y}$ and $m(\beta)$ the number of changepoints of the optimal segmentation result obtained using β in problem 3.7, the theorem writes as follows:

Theorem 3.3.1. *Let $\beta_0 < \beta_1$, 3 cases are possible to uncover new penalty values:*

1. *If $m(\beta_0) = m(\beta_1)$ then $m(\beta) = m(\beta_0)$ for all $\beta \in [\beta_0, \beta_1]$*
2. *If $m(\beta_0) = m(\beta_1) + 1$ then $m(\beta) = m(\beta_0)$ for all $\beta \in [\beta_0, \beta_{int}[$ and $m(\beta) = m(\beta_1)$ for all $\beta \in [\beta_{int}, \beta_1]$ with:*
$$\beta_{int} = \frac{U_{m(\beta_1)}(y_{1:n}) - U_{m(\beta_0)}(y_{1:n})}{m(\beta_0) - m(\beta_1)} \tag{3.10}$$
3. *If $m(\beta_0) > m(\beta_1) + 1$ and $m(\beta_{int}) = m(\beta_1)$ where β_{int} is defined in 3.10, then $m(\beta) = m(\beta_0)$ if $\beta \in [\beta_0, \beta_{int}[$ and $m(\beta) = m(\beta_1)$ if $\beta \in [\beta_{int}, \beta_1]$*

As stated in Haynes et al. (2017), the theoritical upper bound for the number of times PELT has to run to find all segmentations possible with $\beta \in [\beta_{min}, \beta_{max}]$ is given by $m(\beta_{min}) - m(\beta_{max}) + 1$. Given the sizes of signal $\boldsymbol{y}$, CROPS is still more cost-effective than the optimal partitionning method and provides the compromise we were looking for. Note that an estimation of the number of changes can be obtained by combining the results of Algorithm 4 with the elbow method. For each β uncovered by CROPS, we have the number of changes $m(\beta)$ and the cost $\mathcal{C}_\beta(\boldsymbol{y}_{1:n})$. We can run Algorithm 2 to estimate an optimal number of changes. Keeping the results of other segmentations found by CROPS for exploratory purposes.

3.4. Change-points detection in environmental data

Widely 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), and, in particular, for environmental pollution monitoring Costa et al. (2016), change-point detection is a reference technique for time series segmentation. However, the characteristics of concentration data are

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

the censorship and the spatial heterogeneity. We are then looking for studies on the combination on these aspects.

How to handle left-censored and right-skewed data is therefore one aspect to consider when modelling environmental data. A rich literature has been developed on this topic during the last thirty years, and may be roughly divided into three categories of approaches: substitution methods (censored data is imputed using some values chosen *a priori* or via a generative model), parametric methods (maximum likelihood estimates are computed under the hypothesis that the data comes from some log-normal, Weibull, Gamma, exponential, or other log-logistic distribution), and non-parametric methods (Kaplan-Meier or hazard function estimates). Detailed reviews of the various approaches are available for instance in European Food Safety Authority (2010); Hewett & Ganser (2007); Mitra & Kundu (2008); Canales et al. (2018); Antweiler & Taylor (2008); Gillespie et al. (2010); Shoari & Dubé (2018). These studies are providing useful tools to analyse left-censored data but they do not treat change-points detection methods.

The second aspect to consider is spatio-temporal heterogeneity. Air pollution data has received, for instance, a great deal of attention, and several modelling approaches have been proposed in the literature. Some are based on temporal regression models combined with kriging 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 monitoring data is not normally distributed, and is usually left-censored. In the specific field of pesticide concentration monitoring, several recent papers address the spatio-temporal issue from an exploratory point of view see for instance Ccanccapa et al. (2016); Figueiredo et al. (2021); Aznar et al. (2017).

Other methods were found using search associating key words such as change-point, pesticide,

left-censored or environment. The resulting applications domain being very specific, the span of the methods developped in these papers is very large. Several phenomenons are investigated such as resistance appearance to a substance de Solla et al. (2010), the evolution mortality rates and suicide by ingestion of pesticides Ko et al. (2017), the evolution of the exposition of animal populations to pesticide Menger et al. (2022) or impacts of policies on air quality FOMBY & LIN (2006), evolution of annual streamflows Ryberg et al. (2020). However, several common points can be distinguished:

- Many papers focus on changse in trend. Given the data characteristics described in 2.3, trend detection does not seem of interest in our case.
- Indicators on which change point detection is made often have a yearly temporal resolution. We are interested in a finer level of resolution.
- We are also interested in censorship and none of these studies focus on censored indicators. The aggregation of temporal information into yearly informations or rates could explain the absence of censorship in the data.

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 that algorithm to detect homogeneous temporal periods of times 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. 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, see Chapter 5 for an illustration on pesticide concentration data modeling.

- We use the PELT search method to obtain optimal solution to the change-point detection problem. Several penalty values β are explored with the CROPS algorithm. The elbow method is applied when it is necessary to estimate an optimal number of change-points.

We first describe the model integrating the censorship information in Section 4.1. However, we don't known how much the censorship can affect a parametric change-point model. We provde a study of censoring effects in Section 4.2. Futhermore, we need to devise a estimation procedure that is adapted to the observations of pesticide concentrations. The question sums up to detecting breaks in all dimensions of the parameters of Q or not. We devise our estimation scheme in Section 4.3. Finally we test our method with a change-point method adapted for censored data in Section 4.4.

4.1. Generic model for censored data

We present here the underlying parametric model we are using. We condider $\boldsymbol{c} = c_1, \dots, c_n$ that are realizations of independant random variables $C_1, \dots, C_n$. The variables C_i are recorded sequentially, and the recording times are not necessarily equidistant. Thus, the indices in Y_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 index $0 = \tau_0^* < \tau_1^* < ... < \tau_k^* < ... < \tau_{K^*}^* < \tau_{K^*+1}^* = n$. Moreover, on the k-th segment, $C_{\tau_{k-1}^*+1:\tau_k^*}$ follows 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:

$$c_t \sim f(.; \theta_k^*) \mathbb{1}_{\tau_k^*+1 \leq t \leq \tau_{k+1}^*},$$

with f being the density function of distribution Q.
The observations are subject to censorship. We focus on left-censorship because it is adapted for modeling concentration data but similar models can be created for right censorship or a mix of both. To each c_i is associated a known censoring threshold a_i. The resulting censored observations are defined by:

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

Since the C_i are independant and the a_i are known deterministic values, the Y_i are independant as well. The observations of Y_i are denoted y_i. Noting F the cumulative distribution function

(cdf) of Q, we can write the cost function associted to segment $y_{u:v}$ as:

$$W(y_{u:v}) = -\sup_{\theta\in\Theta}\{\sum_{i=u}^{v}\log(F(y_i,\theta))\mathbb{1}_{y_i=a_i} + \sum_{i=u}^{v}\log(f(y_i,\theta))\mathbb{1}_{y_i>a_i}\} \tag{4.2}$$

Note that if one needs to integrate right censorship 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 $S(t) = 1 - F(t)$. For a segmentation $\mathcal{T} = \{\tau_1, ..., \tau_K\}$, the penalised cost is given by:

$$\mathcal{C}(\boldsymbol{y}, \mathcal{T}) = \sum_{i=0}^{K} W(y_{\tau_i+1:\tau_{i+1}}) + KP\beta, \tag{4.3}$$

where P is the dimension of the parameters vector in the distribution Q. Gathering 4.2 and 4.3, this resulting estimator can be expressed as:

$$(\widehat{\mathcal{T}}, \widehat{\boldsymbol{\theta}}) = \arg\min_{\mathcal{T},\boldsymbol{\theta}} \left(-\sum_{i=0}^{|\mathcal{T}|}\left\{\sum_{j=\tau_i+1}^{\tau_{i+1}}\log(F(y_j,\theta))\mathbb{1}_{y_j=a_j} + \sum_{j=\tau_i+1}^{\tau_{i+1}}\log(f(y_j,\theta))\mathbb{1}_{y_j>a_j}\right\} + \beta KP \right) \tag{4.4}$$

Under this configuration, we know that the estimators computed 4.4 have satisfying convergence properties (Lavielle, 1999). A small proof is provided in Appendix A.1.

4.2. Censorship effects

We are interested in the effects of censorship on two main aspects:

- The estimation of the parameters $\boldsymbol{\theta}$ of Q on a fixed segment.
- The implication it can have on the search method PELT and how to adapt it if this is the case.

Illustrating examples are provided in this section with Q set as the exponential distribution.

4.2.1. On the parameter estimation for one segment

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. The cost functions need to be twice differentiable in θ. We search for the zeros of the first derivate of the cost function in order to find a global minimum.
Checking that the second derivative is strictly positive in presence of censored data, thus guaranteeing the unicity of the maximum likelihood estimate, can prove to be a difficult task as it is not always the case for all distribution Q. We provide an example of such a case in Appendix A.2 where we prove the existence of a global minimum without having the global convexity of the first derivate. This implies that a careful initialization of the Newton-Raphson

method to obtain convergence. Appendix A.2 also provide experiments on the initialization of the Newton-Raphson method.
Specifically, the case where all data in the segment $y_{u:v}$ are censored is problematic. Looking at the analytical likelihood formula, we find that the θ realising the minimum of the cost function is still unique, but tends toward infinity. In this case the cost tends to zero. We illustrate in Figure 4.1 with Q set as an exponential distribution.
For a given segment $y_{u:v}$ where all observations are censored and under a censorship threshold a we have that:

$$W(y_{u:v}) = -\sup_{\theta \in \Theta}(v-u)\log(1-\exp(-\theta a)) \tag{4.5}$$

This cost is always positive and decreasing to 0 when θ goes to infinity.

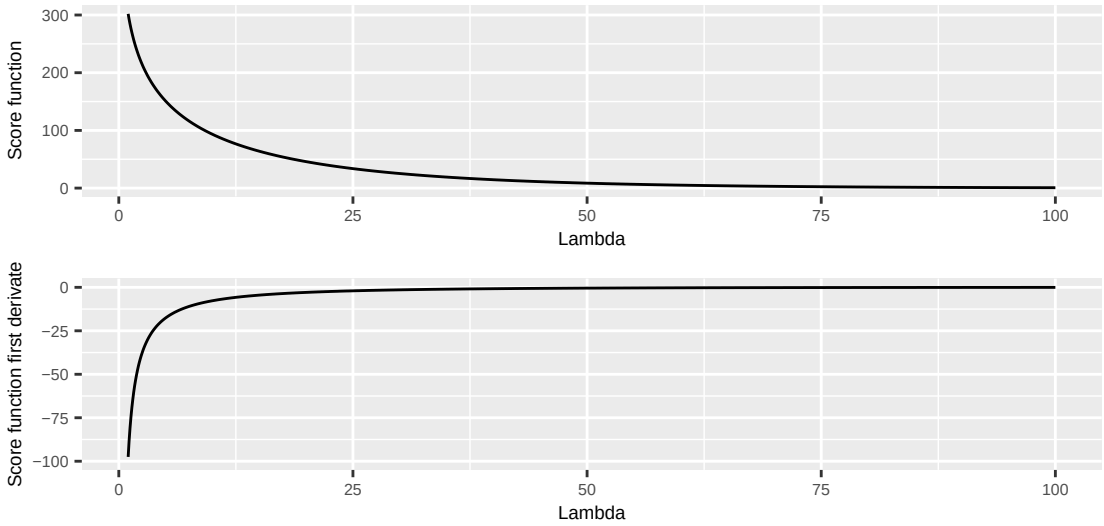

Figure 4.1: Plot of the cost function values against θ values when all observations are censored. It is represented for an exponential distribution. The sample consists in 100 values of censored observations to a threshold $a = 0.05$.

We show in the next part why it could be problematic to have fully censored segments in the search method and how we chose to deal with it.

4.2.2. On the detection point method

The case of fully censored segment are questionning the identifiability of the change-point detection method. The usual assumption (Lavielle, 1999) for the segment parameters is the following:

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 $k = 0, ..., K^\star$.

We have seen in 4.5 that, for the exponential distribution example, the optimal θ tends to infinity. If that is the case Θ is not compact. To solve this problem, we impose an upper bound

θ_{max} on the possible values of θ. Thus, we have that $\theta \in [0, \theta_{max}]$ which is a compact part of $\mathbb{R}$.
A new problem arises from this modeling 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 identification 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 identification will be possible. In order to avoid such problems, the value of θ_{max} is set according to the worst censoring case scenario possible. More precisely, we assume no change-point occured 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:

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

with α being a desired percentage of censorship.
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 $\theta_0^* = 1$ for $y_{1:100}$ and $\theta_1^* = 4$ for $y_{101:200}$. The censoring level is set to the median of $\boldsymbol{y}$ so that 50% of the signal is censored. We illustrates $\boldsymbol{y}$ in Figure 4.2. We set θ_{max} choosing with $\alpha = 95\%$. We have that $\theta_{max} = \frac{-\log(1-\alpha^{1/n})}{\min(\boldsymbol{y})}$. In our numerical example, $\theta_{max} = 24.68$, which is greater than θ_0 and θ_1.

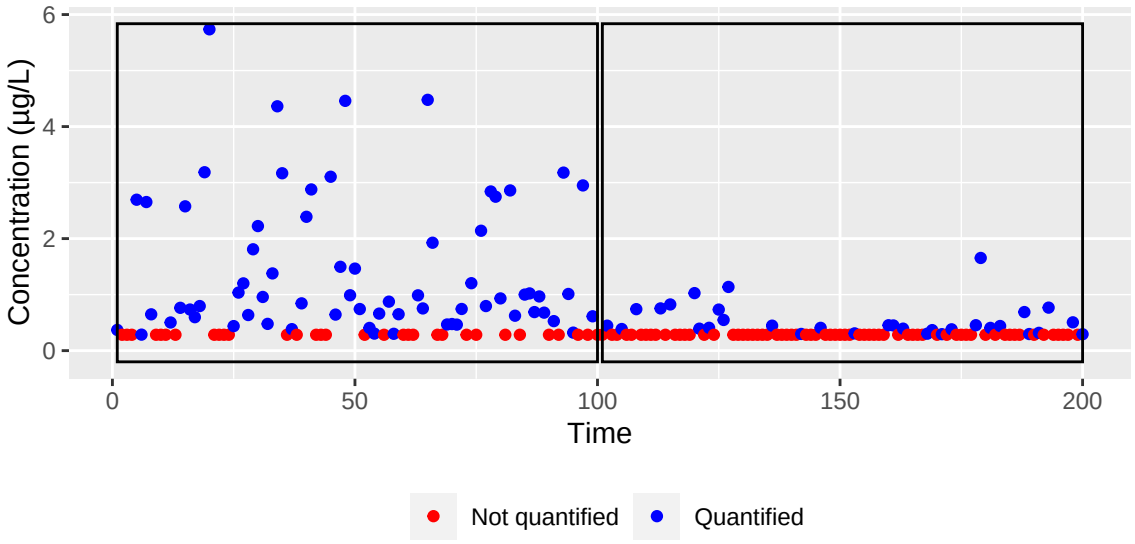

Figure 4.2: Example of simulated signal $\boldsymbol{y}$ distributed according exponential distributions. The two segments are drawned with black rectangles. $\theta_0^* = 1$ in the left segment, $\theta_1^* = 4$ in the right one and the censoring threshold $a = 0.28$ in both segments

Censoring in the data creates identification problems that we chose to deal by adding a new regularisation parameter θ_{max}. Note that other ways to tackle this problem are possible. The modification of the pruning rule 3.9 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 censored segment $y_{u:v}$.

4.3. Multi-parameter estimation

We discuss here the procedure for the parameters estimates proposed in 4.4. When $\theta^* \in \mathbb{R}^P$ with $P > 1$, several optimization strategies are possible. 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 parameters vector of each segment k.
- $\boldsymbol{\theta}_{k,.} = (\theta_{k,1}, \dots, \theta_{k,P})$ is the parameters vector of the k-th segment.
- $\theta_{k,m}$ is the m-th dimension of the parameters vector of the k-th segment.

A common strategy is to opt for simultaneous changes detection in all P parameters simultaneously. This implies detecting changes in different properties of $\boldsymbol{y}$. An example with Q set as a Weibull distribution is provided in Figure 4.3. In practice, it depends on what parameters are optimized in the cost function. The optimization can be petformed with numerical methods (such as Newton-Raphson) on all dimensions of θ simultaneously. For the k-th segment, let $\widehat{\theta}_{NR} = (\widehat{\theta}_{k,1}, \dots, \widehat{\theta}_{k,P})$ the estimates of $\theta_{k,.}$ obtained with Newton-Raphson. Using $\widehat{\theta}_{NR}$ as the supremum in 4.2 provide a cost function that detects changes occuring in any of the parameters.

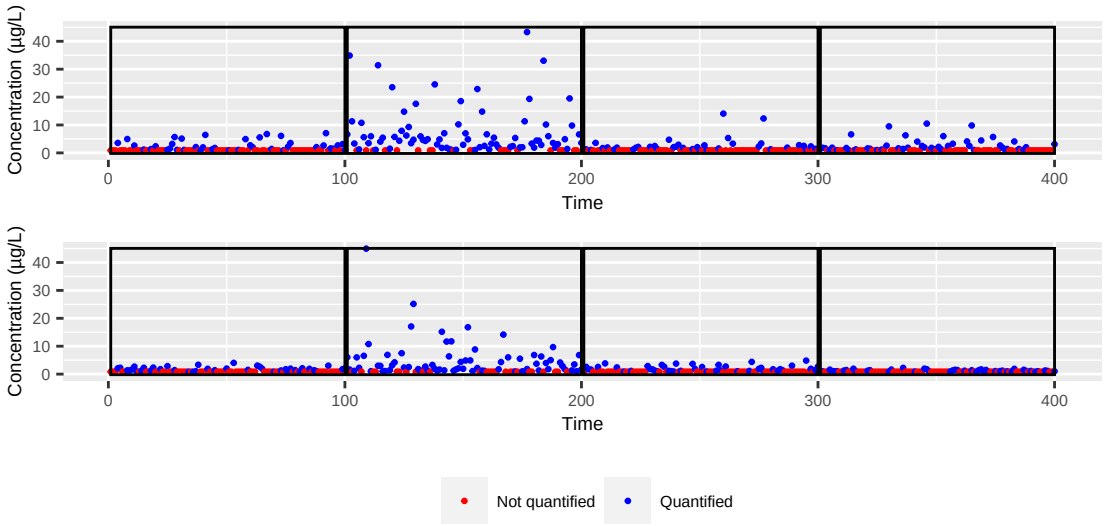

Figure 4.3: Example of simulated signal $\boldsymbol{y}$ distributed according Weibull distributions with (λ, σ) as the scale and shape parameters. The segments are drawned with black rectangles.
Upper signal: The associated parameters to each segment are $\theta^*_{0,.} = (\lambda^*_0, \sigma^*_0) = (1, 1)$, $\theta^*_{1,.} = (\lambda^*_1, \sigma^*_1) = (3, 0.7)$, $\theta^*_{2,.} = (\lambda^*_2, \sigma^*_2) = (1, 0.7)$, $\theta^*_{3,.} = (\lambda^*_3, \sigma^*_3) = (1, 2)$ and the censoring threshold $a = 0.86$ in all segments.
Lower signal: The associated parameters to each segment are $\theta^*_{0,.} = (\lambda^*_0, \sigma^*) = (1, 0.7)$, $\theta^*_{1,.} = (\lambda^*_1, \sigma^*) = (5, 0.7)$, $\theta^*_{2,.} = (\lambda^*_2, \sigma^*) = (0.7, 0.7)$, $\theta^*_{3,.} = (\lambda^*_3, \sigma^*) = (1, 0.7)$ and the censoring threshold $a = 0.89$ in all segments.

We propose a different estimation strategy. We are interested in models where changes occur only in some dimension $\boldsymbol{\theta}^*_{.,m}$. 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. Figure 4.3 provides an illustration of signal $\boldsymbol{y}$ simulated from these assumptions. The parameters $(\boldsymbol{\theta}^*_{.,m})_{m\in\overline{\mathcal{M}}}$ are supposed to be fixed throughout the signal $\boldsymbol{y}$. In that setting, the estimation procedure changes. We can rewrite 4.4 as:

$$(\widehat{\mathcal{T}}, \widehat{\theta}_{.,m\in\mathcal{M}}, \widehat{\theta}_{.,m\in\overline{\mathcal{M}}}) = \arg\min_{\mathcal{T},\boldsymbol{\theta}_{.,m\in\mathcal{M}}} \left[\arg\min_{\theta_{.,m\in\overline{\mathcal{M}}}} \left\{ - \sum_{i=0}^{|\mathcal{T}|} \left(\sum_{j=\tau_i+1}^{\tau_{i+1}} \log(F(y_j,\theta))\mathbb{1}_{y_j=a_j} + \sum_{j=\tau_i+1}^{\tau_{i+1}} \log(f(y_j,\theta))\mathbb{1}_{y_j>a_j} \right) \right\} + \beta KP \right] \tag{4.6}$$

From 4.6, we can design a two-step iterative estimation strategy. The two steps of each iteration are:

1. Compute the MLE $\widehat{\boldsymbol{\theta}}_{.,m\in\overline{\mathcal{M}}}$ with fixed $\widehat{\mathcal{T}}$ and $\widehat{\theta}_{.,m\in\mathcal{M}}$ with any optimization method that handles censored data. We use `R` package developped in Delignette-Muller & Dutang (2015) in our procedure.

2. Run the change-point procedure to estimate $\widehat{\mathcal{T}}$ and $\widehat{\theta}_{.,m\in\mathcal{M}}$ using the values of $\widehat{\boldsymbol{\theta}}_{.,m\in\overline{\mathcal{M}}}$.

In the second step of the iteration, the estimated segmentation $\widehat{\mathcal{T}}$ and the fitted parameters on its segments $\widehat{\theta}_{.,m\in\mathcal{M}}$ are obtained by applying the PELT procedure (Killick et al., 2012). PELT is run several times on a penalty grid $(\beta_0, \ldots, \beta_q, \ldots, \beta_B)$ where $\beta_0 < \cdots < \beta_q < \cdots < \beta_B$ are evenly spaced. We obtain B segmentations of $\boldsymbol{y}$. Eventually, the optimal penalty value for this step is selected using an elbow rule heuristic as proposed in Section 3.2.
The initialization is an important part of this procedure. We would like the initial estimate of the fixed parameters to be close to $\boldsymbol{\theta}^*_{.,m\in\overline{\mathcal{M}}}$ to ensure the convergence of the procedure. In order to do so, we intialise $\widehat{\boldsymbol{\theta}}$ assuming no change-point occured in $\boldsymbol{y}$. More formally, we compute:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\theta} \left\{ - \left(\sum_{j=1}^{n} \log(F(y_j,\theta))\mathbb{1}_{y_j=a_j} + \sum_{j=1}^{n} \log(f(y_j,\theta))\mathbb{1}_{y_j>a_j} \right) \right\} \tag{4.7}$$

We proceed using the MLE estimator for all parameters in $\boldsymbol{\theta}$, which implies using iterative methods again as stated in Cohen (1965). We choose the initial values as the $\widehat{\boldsymbol{\theta}}_{.,m}$ such that $m \in \overline{\mathcal{M}}$. It can be noted straight away that the value of $\widehat{\boldsymbol{\theta}}_{.,m\in\mathcal{M}}$ will be discarded since it was computed from a model that goes in direct contradiction with the assumption in 4.1 that they were change-points in the signal $\boldsymbol{y}$. We show the efficiency of this procedure in Appendix A.3. We use this procedure estimation in Chapter 5. We justify this choice with the observation of real data. We compare the efficiency of our method a state-of-the-art in the next section.

4.4. Simulation study

We test our method on simulated data and compare it with the *Multrank* method Lung-Yut-Fong et al. (2015). This section aims at two objectives. First, we want to calibrate the parameters of our method to ensure good efficiency. Then, we want to test if that calibration really leads to better change-point results.

4.4.1. Calibration of the minimal segment length

In section 2.2, the PELT Algorithm 3 introduces a minimal segment length n_{min}. We lead some simulation tests to calibrate its value. We need to identify n_{min} so that the cost function defined in 4.2 has sufficient data to make the difference between segments on which the parameters differ.

This task can be viewed as a classification one. We want to know if our method is able to classify correctly signals that have a change point or not. Since we use a parametric inference (see Section 4.1), we use the log likelihood ratio test to assess the presence of a change-point. This is a very common test used in change-point that can be found in. The statistics of the likelihood ratio test are calculated on several simulated signals and and the ROC curves (Fawcett, 2006) are derived from it. We compute the corresponding areas under the curve (AUC) to assess the ability to detect change-points of our 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. We will use this comparison to calibrate our method performance.

The simulation procedure for the calibration of n_{min} is the following. We use the Weibull distribution for the distribution Q. We test several configurations with different censoring levels $\alpha = (5\%, 25\%, 50\%, 75\%, 95\%)$. It is a global censoring level, meaning that for a given signal $\boldsymbol{y}$, $\alpha\%$ of the data is censored. The detection methods are run with different values of minimal segment size $n_{min} = (5, 10, 25, 50, 75)$. For each configuration, we simulate $M = 1000$ signals of size $n = 200$ among which half of them present a change point in position 100. The parameters of signals with a change-point are $(\sigma = 0.5, \lambda_0^* = 1)$ on the first segment and $(\sigma = 0.5, \lambda_1^* = 3)$ on the second. The parameters associated to signals with no change-point are $(\sigma = 0.5, \lambda = 1)$. We compute the AUC on these M signals. The results are illustrated in Figure 4.4.

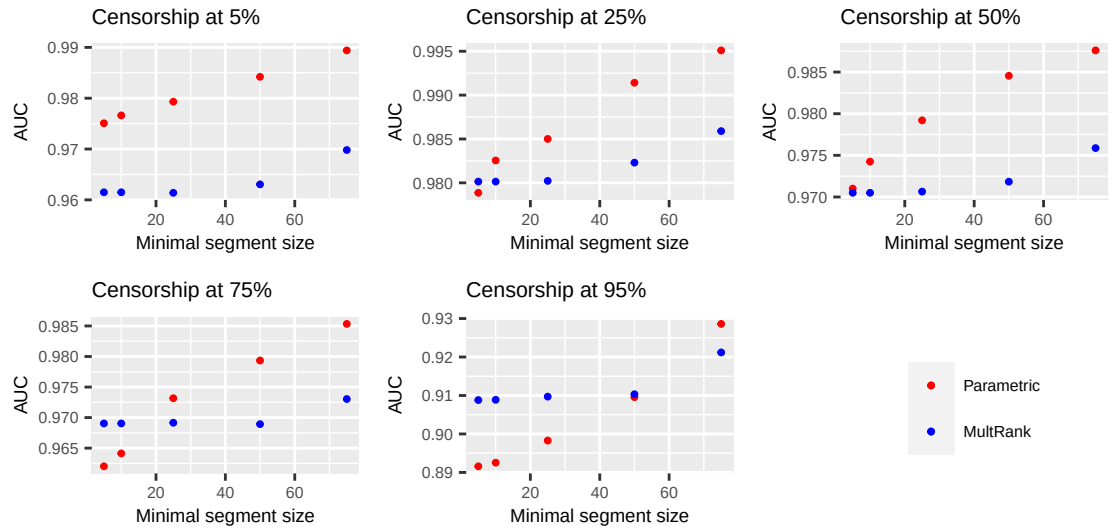

Figure 4.4: Choice of the minimal segment length: simulation results. Our method performance is Illustrated with the red dots, the *Multrank* method is drawn in blue.

From the results of Figure 4.4, 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 levels increase, the more data is needed in the minimal segment length of the parametric method to outclass the non parametric method.

From this study, we can conclude that when working with real data, we should get the censoring level information information in order to calibrate the minimal segment length of the parametric method.

4.4.2. Comparison with a non parametric method

We want to compare the performance of the change-point detection with the Multrank method developped 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 $N = 100$ samples $(x_1, ..., x_n)$ of size $n = 400$ following a left-censored Weibull distribution with $\alpha\%$ of censored data. We made tests for the different censorship rates $\alpha = (25, 50, 75, 95)$. The shape parameter of the Weibull distribution is assumed to be known and set to $\sigma = 0.5$. The scaling parameters $\boldsymbol{\lambda}^\star$ have $K^\star = 4$ breaks at positions $p_1^\star = 80$, $p_2^\star = 160$, $p_3^\star = 240$ and $p_4^\star = 320$ and take the values $\boldsymbol{\lambda}^\star = (\lambda_1^\star = 1, \lambda_2^\star = 4, \lambda_3^\star = 0.5, \lambda_4^\star = 5, \lambda_5^\star = 1)$. An example of a sample simulated in this way is shown in Figure 4.5.
2. For each of the N samples, we perform the parametric change-point detection and the Multrank methods. For each sample, we obtain the estimated number of breaks $\hat{K}_{param}$ and $\hat{K}_{multrank}$ and their position $(\hat{p}_{k,param})_{k=1}^{\hat{K}_{param}}$ (respectively $(\hat{p}_{k,multrank})_{k=1}^{\hat{K}_{multrank}}$).
3. for both methods, we count the number of samples among the N for which the correct number of breaks has been estimated (e.g. $\hat{K}_{param} = K^\star$). Also, for each of the samples for which the estimate of $K^\star$ is correct, we make an histogram of the change-points position.

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

- For the parametric method: we use the algorithm CROPS, algorithm to scan a continuous range of penalty values $[\beta_{min}, \beta_{max}]$. We obtain a set of B values $(\hat{\beta}_1, ..., \beta_B)$ and the optimal segmentations associated with these penalty values. We then plot the cost of the segmentations as a function of the number of breaks. We choose the optimal penalty using a elbow heuristic. This procedure is described in Haynes et al. (2014). 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 using the optimal segmentation search method presented in Algorithm 1 for k breaks, where k ranges from 1 to K_{max}. For each of these segmentations, we can compute the cost of the segmentations using the cost function explicited in 3.5. As in the parametric method, we represent the costs as a function of k, and we determine the number of estimated breaks by an elbow heuristic. Here, K_{max} is fixed at $2 * K^\star = 8$.

The results of the simulations are shown in Table 4.1 and in Figure 4.6. It can be seen that in the ideal scenario, where the data are indeed distributed according to a left-censored Weibull distribution, the parametric method performs better both in detecting the correct number of breaks and in accurately estimating their position. However, this performance decreases as the censoring rate increases.

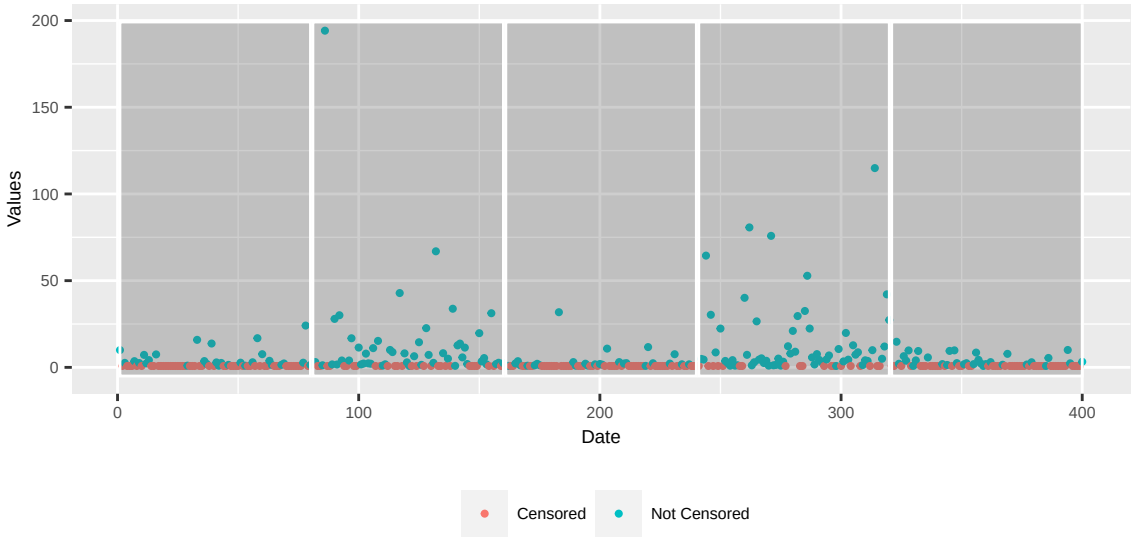

Figure 4.5: Example of simulated signal with $(\lambda_1 = 1, \lambda_2 = 4, \lambda_3 = 0.5, \lambda_4 = 5, \lambda_5 = 1)$, $\sigma = 0.5$, $n = 400$, $K = 4$, $(p_1 = 80, p_2 = 160, p_3 = 240, p_4 = 320)$ and $\alpha = 50\%$.

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

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

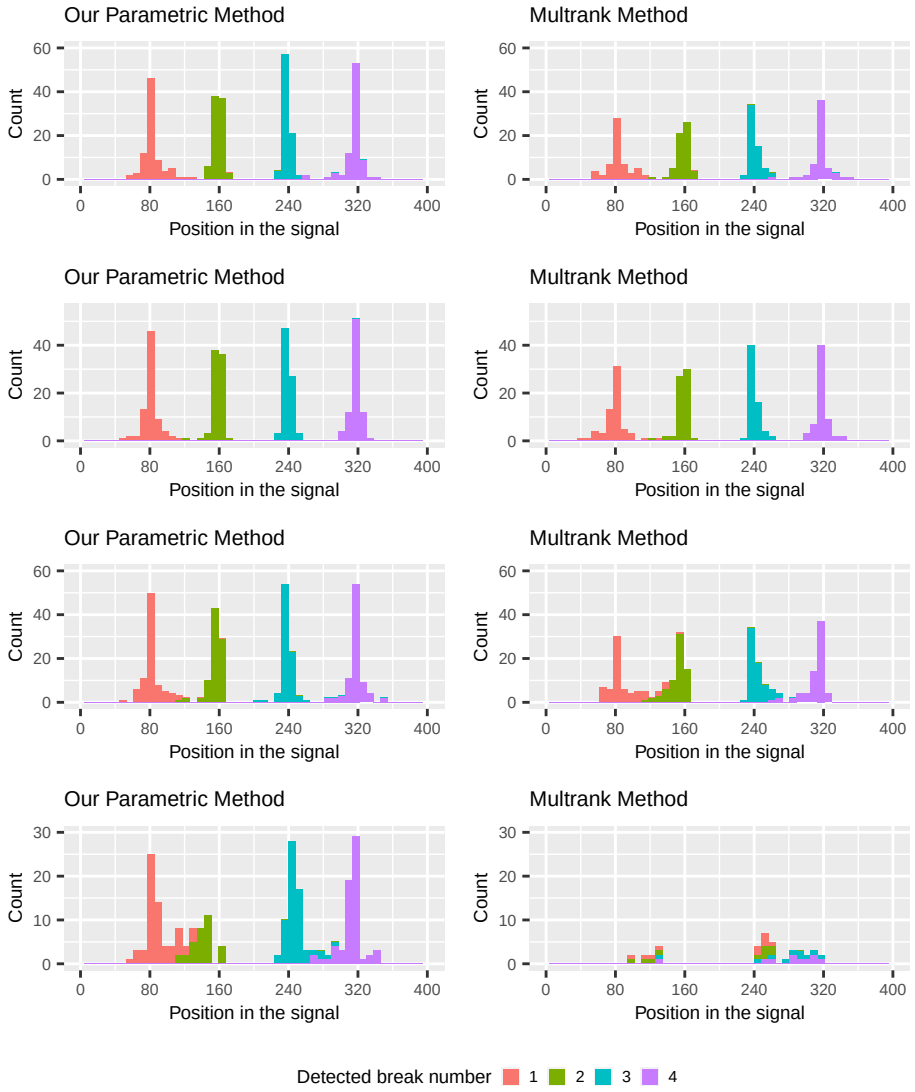

Figure 4.6: Precision of the estimated change-points for both methods.

5. Spatio-temporal analysis of concentration data

Contents

This chapter focuses on the spatial dimension of concentration data. Spatial clustering and anomaly detection are the key elements that will allow to detect anomalous geographical areas where abnormal concentration values were collected. This chapter is fully articulated with the temporal change-point detection. We illustrate the whole procedure on real dataset of concentration of a substance monitored in surface waters.

If one focuses specifically on temporal heterogeneity, the approach we use to deal with it is to use a change-point based segmentation developped in Chapter 4. We assume the data 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. While the literature on change-point detection is abundant (see Chapter 3.4), 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).

In this chapter, we tackle the issue of pesticide concentration monitoring, and introduces a new methodology which integrates both the specific left-censored distribution of the data, and the spatio-temporal context. The main goal is to identify contextual anomalies, both from a temporal and a spatial point of view. The proposed method builds on a parametric model for left-censored and right-skewed distributions, and combines it with a change-point detection step and a clustering step.

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

Clustering is then used for modelling the expected spatial homogeneity while integrating geographical constraints such as river networks, wind directions, etc. Indeed, as geological, terrain and climatic characteristics of an area can influence the dispersion of a chemical substance and on its potential use in the case of e.g. a pesticide, concentrations are expected to be somewhat correlated in small scale regions that are homogeneous in terms of influencing characteristics. Especially in the application presented here, which relates to the investigation of pollutants in surface waters, it is interesting to take into account the hydrographic structure of the region as in e.g. 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. This hypothesis is accounted for by building clusters of measuring stations according to their proximities measured via the hydrographic network.

Conditionally to the temporal segment detected by the change-point procedure, and to the spatial cluster detected by the clustering procedure, one may analyse the data and identify contextual anomalies.

The rest of the chapter is organised as follows: in Section 5.1, the generative model assumed for environmental pesticide monitoring data is described; the proposed method for estimating and handling this model from observed data is detailed in Section 5.2; a detailed example on data collected by French authorities in Val de Loire region is fully illustrated in Sections 5.3 and 5.4.

5.1. Data collection procedure and associated generative model

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

5.1.1. Monitoring stations network

We consider a network of monitoring stations used to collect concentration measurements at irregularly sampled instants. The stations are represented by an undirected graph $G = (V, E)$, which vertices $V = (v_i)_{1 \leq i \leq N}$ are the monitoring stations and which weighted edges E are links between stations that are directly comparable. The aim of the graph is to represent expert knowledge about expected measurement homogeneity. When two stations are connected in G, their measurements can be compared directly: a small edge weight assumes simultaneous measurements to be close, while a large one allows for significant differences. Shortest paths in the graph can be used to compare stations that are not directly connected, using the total weight of the paths to measure non homogeneity. This approach is inspired by methods developed for signal processing on graphs 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 5.2.1). Other modelling approaches may use a different graph considering for instance dominant wind directions relevant for air diffusion of pollutants.

This graph is not necessarily fully connected, there can be P non connected components that we will denote $(\mathcal{K}_1, ..., \mathcal{K}_P)$.

5.1.2. Data collection

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

Summarising the above notations and hypotheses, a data set sampled from the stations network

is given by a collection of measurements and associated quantification limits, and denoted

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

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

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

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

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

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

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

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

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

- there are $L^* > 0$ change-points producing $L^* + 1$ stationary intervals defined by
$$0 = \eta_0^* < \eta_1^* < \ldots < \eta_{L^*}^* < \eta_{L^*+1}^* = K;$$
- the observations $(\overline{y}_k)_{1 \leq k \leq K}$ are realisations of K independent random variables $(\overline{Y}_k)_{1 \leq k \leq K}$;
- when $k \in [\eta_{l-1}^* + 1, \eta_l^*]$, $\overline{Y}_k$ is distributed according to a left-censored parametric Weibull distribution with interval dependent parameters λ_l^* and a left-censoring threshold $\overline{q}_k$, which is a known constant. The shape parameter σ^* is supposed unknown and fixed throughout the whole signal.

Several remarks must be pointed out at this point. First, notice that the model only accounts for the concentrations $\overline{y}_k$ but not for the instants and the quantification limits which are supposed deterministic quantities. The second remark is that the $(\eta_l)_{l=1}^{L^*+1}$ define the **contextual** aspect for the anomaly detection step implemented in 5.2.2.

5.2. Methods

5.2.1. Spatial clustering

In any of stationary intervals identified in the previous step, the measurements are assumed to be consistent with the homogeneity assumptions represented by the graph $G = (V, E)$. A natural way of assessing the actual regularity of the measurements would be to use graph signal processing techniques see e.g. Ortega et al. (2018); Shuman et al. (2013). However the irregular, unaligned, sparse and censored nature of the measurements at each station, prevents the use of such methods. The measurements are also incompatible with techniques designed to detect anomalous clusters in a graph see for instance Arias-Castro et al. (2011).
To circumvent this problem, we propose to leverage the graphical representation to build spatial aggregates and to assess homogeneity at this aggregated level. This corresponds to clustering the stations using the graph structure. Nodes of each connected component $(\mathcal{K}_1, ..., \mathcal{K}_P)$ of the graph $G = (V, E)$ are clustered using a Ward hierarchical clustering method implemented on the shortest path distance computed from the edge weights.
The goal is to successfully create a global partition in M cluster of stations in the presence of P non connected components in the graph. This raises the question of how to dispatch these M clusters among the non connected components. We have developed two methods. The first one proceed in a greedy way, the second is based on dynamic programming. Both ot them are based the standard definition of inertia given for the clustering $\mathcal{P} = (C_1, \ldots, C_M)$ by

$$W(\mathcal{P}) = \sum_{m=1}^{M} \frac{1}{|C_m|} \sum_{v_i, v_j \in C_k} d_{ij}^2, \tag{5.4}$$

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

1. **The greedy clustering method:** the initial global clustering of V is obtained by assigning all vertices in a connected component to the same cluster. Subsequent levels of the global hierarchy are obtained by replacing the clusters of a connected component by the next refined level of the local hierarchy. At each step of the refinement, we select the component that reduce the most the inertia of the clustering 5.4.

2. **A clustering method based on dynamic programming:** this approach is derived from Hébrail et al. (2010). This paper shows that is possible to create a partition of the stations graph into P components and to perform a segmentation of each the P components using a total number M of segments. The M segments are distributed among the P components in an optimal way using dynamic programming. Our context is a little bit simpler than Hébrail et al. (2010) since the P components are already known and doesn't have to be estimated.

The algorithms for both methods are provided in Appendix B.1. To select the final clustering in the hierarchy, we use the same elbow method heuristic as in Algotiyhm 2. This time, the

inertia of the clustering is plotted against the corresponding number of clusters M. We look for the number of breaks M^* that minimizes the sums of squares of two linear models respectively fitted on the $M \geq M^*$ and the $M \leq M^*$.
Notice that we rely on a simple graph clustering approach for two main reasons. Firstly, we do not expect graphs of monitoring stations to exhibit the specific characteristics of complex networks (such as very high degree vertices, small diameter, etc. see e.g. 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 the spatial clustering is independent from the temporal context $[\widehat{\eta}_l, \widehat{\eta}_{l+1}]$. The clustering is performed on the graph $G = (V, E)$ composed of all stations available in the data.

5.2.2. Anomaly detection

Two types of anomalous clusters are targeted: either clusters with anomalous stations, or wholly anomalous clusters. Clusters containing anomalous stations are detected by studying the homogeneity of the measurements provided by the stations in a given spatial cluster. Anomalous clusters of stations are detected by simply pooling all measurements of each cluster to estimate the local use of the substance and detect large rates. We derive in this section two anomaly scores covering those cases.
For the first case, we need to assess the homogeneity of the measurements of the stations in a spatial cluster for a stationary time interval. As pointed out previously, the number of measurements provided by a single station is usually quite small, especially when we consider a single stationary interval. As a consequence classical distances between empirical distributions are not appropriate, mainly because the measurements of two stations do not have any value in common. Then the Kolmogorov-Smirnov statistics will be essentially driven by the number of observed values rather than the actual values, while other quantities, such as the Jensen-Shannon divergence, cannot be properly estimated (see appendix B.2). For this reason, we propose to use the Wasserstein w_1 distance Villani (2009) adapted for left censored variables. For two discrete distributions on $\mathbb{R}$, it is expressed as the L^1-distance between their cumulative distribution functions and is therefore simple to compute.
The measurement homogeneity of the clusters obtained in Section 5.2.1 is therefore defined as the mean within cluster empirical Wasserstein average distance of a station measurements to the others. Denoting C_m the m-th cluster and $|C_k|$ the number of stations present in C_m, $w_1(\mathbf{y}_i, \mathbf{y}_j)$ the empirical 1-Wasserstein distance between the data of stations v_i and v_j, this quantity is expressed as

$$\bar{W}_k = \frac{1}{|C_m|(|C_m|-1)} \sum_{1 \leq j \leq |C_m|} \sum_{1 \leq i \leq |C_m|, i \neq j} w_1(\mathbf{y}_i, \mathbf{y}_j). \tag{5.5}$$

The second type of potentially anomalous clusters are simply associated to the presence of quantified measurements and high values of concentration. Thus we estimate for each spatial

cluster C_m the parameters of distribution Q (see Section 5.1.3) on the pooled measurements obtained from all the stations of the cluster during the chosen stationary interval. From those parameters, we compute a statistics, denoted $\bar{I}_m$, used as a proxy for the intensity of the measurements (see Section 5.4.3 for an example). Hence we consider a low concentration to be the normal case, but we do not define a threshold between normal clusters and abnormal ones. Each cluster C_m is therefore characterised by two values $(\bar{W}_m, \bar{I}_m)$. To select potentially anomalous clusters, we use a multi-objective optimisation approach, considering that both characteristics are equally interesting. Following Kießling (2002), we say that $X_k = (\bar{W}_m, \bar{I}_m)$ is *Pareto dominated by* $X_l = (\bar{W}_l, \bar{I}_l)$, and we write $X_m \prec X_l$ if and only if

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

The level 1 Pareto optimal front is the set of maximal points for $\prec$. Level b with $b > 1$ is defined recursively as the optimal Pareto front computed for the set of points that do not belong to the optimal Pareto front of levels $1, \ldots, b-1$. Therefore clusters in the level 1 Pareto front are remarkable in the sense that there is no other cluster with higher heterogeneity and more extreme measurements. We define these clusters as anomalous. Pareto front and levels are evaluated using the Skyline algorithm Borzsony et al. (2001); Endres et al. (2015).

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 (2022) 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) (2018); Agriculture & Environment Research Unit (AERU) at the University of Hertfordshire (2021).

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[1]) in 2009. Since then, two modifications of the authorisation for use have been put in place, in November 2018 and in November 2019 respectively. Both changes consist in restrictions of use, one imposing specific equipment for application, the other restricting the application schedule in the presence of non-target crops next to the treated area. Motivated by these changes in regulation, the time period chosen for our study spans from January 1, 2007, to April 8, 2022. Moreover, our study focuses on the geographical area of French Centre-Val de Loire region. Indeed, between 2009 and today, the annual mass of prosulfocarb sold in this region exploded, making it rise from the 17th most sold substance in 2009 to the 4th in 2017 (see Figure B.4 in Appendix B.3.2).

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

This region is also characterised by high concentrations of prosulfocarb target crops (such as the Beauce plains) (see Figure B.3 in Appendix B.3.1). Many target crops (cereal crops) are also concentrated in the region. These two elements combined guarantee a significant use of the product in this area. Thus, we expect significant variations in concentration values in this area during this period.

Data about surface water quality in France is available from the French Biodiversity Agency Office français de la biodiversité (n.d.). We collected from the site the data selected above[2]. These choices led to a data set $\mathcal{D}$ comprising 420 monitoring stations that performed 14,203 measurements. Each measurement is described by the monitoring station ID, the sampling date, the quantification limit (LOQ), and the concentration measurement value, if the concentration exceeds the LOQ. In the data used in this work, the LOD is unknown: the left censoring phenomenon corresponds therefore to the LOQ of the measuring stations. When both limits are known, one can adapt the model proposed in Section 5.1.2 to take both of them into account: this would translate into a slightly more complex likelihood as the one derived in Chapter 4 as we need to consider three cases (when the concentration is between 0 and the LOD, when the concentration is between the LOD and the LOQ, and finally when the concentration is observed and larger than the LOQ).

Among the 14,203 recorded measurements during the period of interest, only 14.11% were above the quantification limit. Figure 5.1 shows the distribution of the number of measurements per station: the mean (rounded to the closest integer) and median number of samples collected by each monitoring station are respectively 34 and 19. This illustrates that sampling rates are different across stations, most of them making few measures, and the monitoring process is heterogeneous.

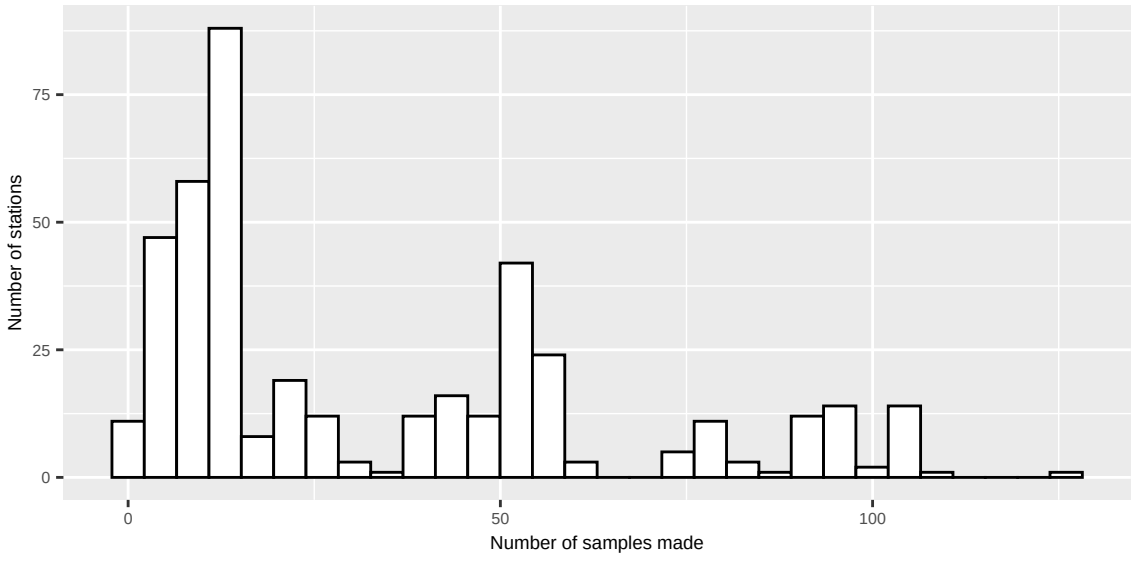

Figure 5.1: Distribution of the number of measurements per station.

[2]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`

The coarse representation $\overline{\mathcal{D}}$ of the monitoring data $\mathcal{D}$ is obtained by computing the maximum daily values across the available stations. This yields the time series illustrated in Figure 5.2. The aggregated series contains 2,150 values, among which 22.51% are quantified.

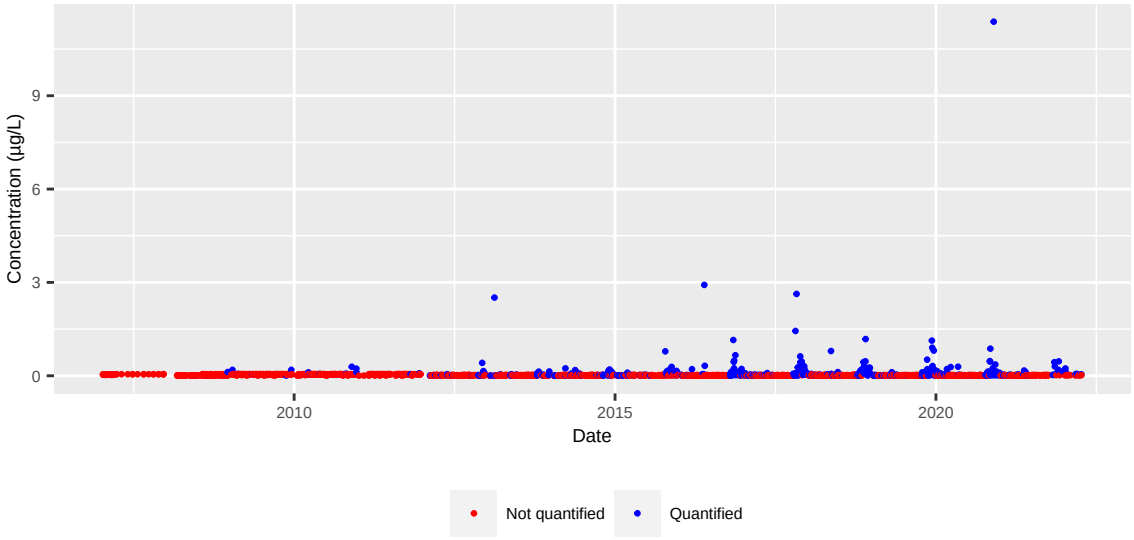

Figure 5.2: Plot of daily maximum concentrations

One may note here that despite the aggregation process, the coarse series remains irregularly sampled, and that for about two thirds of the days included in the studied time span, no measurements were made.

5.3.2. Graphical representation of the station network

The stations network $G = (V, E)$ introduced in Section 5.1.1 is built using the hydrographic map of the Centre-Val de Loire region. Indeed, once the monitoring stations are geo-localized through their GPS coordinates, one still has to compute the edges between them, as well as the associated weights.
For the data at hand, edges are determined using the river network. A database provided by the French National Institute of Geographic and Forest Information (IGN) Institut National de l'Information Géographique et Forestière (2021) contains a fine-grained description of rivers, encoded as sequences of hydrographic sections (or river sections). River sections are segments with constant geographic and hydrographic attributes.
The procedure used for computing the edges in the stations network based on the river network may be summarised as follows:

1. One starts by building a river network $R = (S, H)$, where the vertices S are made of the connecting points between the river sections, and the edges H contain all sections. Each edge is thus naturally weighted by the length (in meters) of the corresponding river section.

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

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

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

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

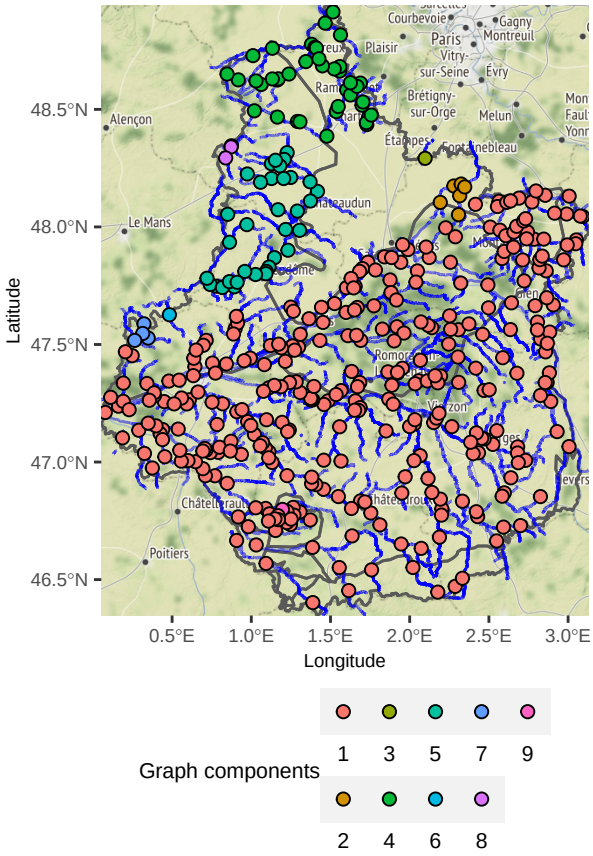

Figure 5.3: Map of the non connex components in the station graph.

5.4. Results

5.4.1. Temporal segmentation

First, the coarse-grained time series $\overline{\mathcal{D}}$ in Figure 5.2 is segmented using the change-point detection procedure described in Section 4.3. We fitted a left censored Weibull distribution in the change-point model with parameters (λ, σ). This was motivated by the observation of the data. We suppose σ fixed in the time series and propose the heuristic 4.3 to estimate it. The results of its estimation are shown in Figure 5.4.

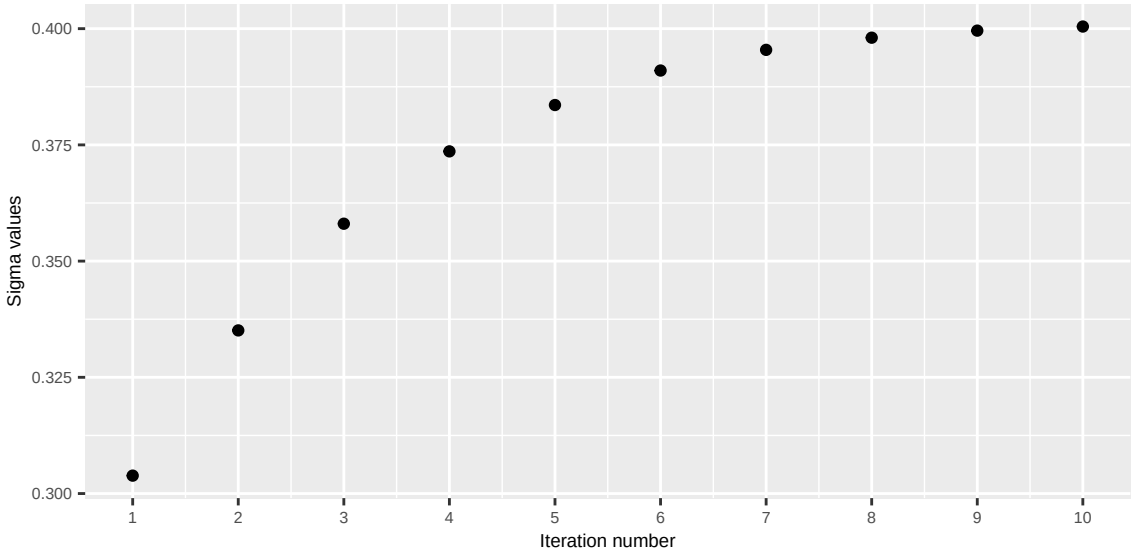

Figure 5.4: Plot of successive $\widehat{\sigma}$ values. We stopped the to iterate when the $|\widehat{\sigma}_b - \widehat{\sigma}_b| \leq 10^{-3}$

The penalty grid $[\frac{\ln K}{4}, 4 \ln K]$ was chosen for the heuristic where K is the number of daily maximum concentrations available, here $K = 2150$.. This interval allowed to obtain various segmentation results with the number of change points varying from 1 to 30. With such kind of range in the results, it ensured a central position of the elbow in the plot of the cost of these segmentations against their respective number of change-points. The precision stopping criterion was set to 10^-3. From the application point of view, the assumption that σ is a fixed parameter throughout the series $\overline{\mathcal{D}}$ corresponds to the hypothesis that the differences in usage and diffusion of the prosulfocarb among the different users is captured by the shape parameter, and should not vary much over time. On the contrary, the overall average usage of prosulfocarb varies, and this dependency is captured by changes in the rate parameter. Let us remark here that the estimated value of the shape parameter is $\hat{\sigma}_{MLE} = 0.4$. This confirms the data has a heavier tail than an exponential distribution (σ=1), and that the assumption of using Weibull distributions for our data is appropriate.
In the change-point detection procedure, the penalty value for the PELT algorithm was calibrated using a large range of values explored according to the CROPS algorithm. The range, inspired by the BIC criterion, was set to $[\frac{\log(K)}{5}, 5 \log(K)]$. Note that when using the BIC penalty in change point detection, the penalty term written becomes : $\beta_K(L+1)D = \frac{D}{2} \log(K)(L+1) = \frac{1}{2} \log(K)(L+1)$. The range chosen allows to screen an interval of penalties containing the BIC penalty.

The penalty calibration procedure resulted in 15 different segmentations, with a number of change-points ranging from 1 to 30. The best segmentation is selected using the elbow method, as illustrated in Figure B.5 in Appendix B.3.3. This amounts to a temporal segmentation with $\hat{L} = 13$ change-points, illustrated in Figure 5.5.

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

According to Figure 5.5, the usage of prosulfocarb in Centre-Val de Loire shows different patterns throughout time. Before 2016, most of the values are not quantified, and there are almost no change-points detected. Starting with 2016, two regimes of pesticide usage appear to emerge, and correspond respectively to the periods of intensive usage of prosulfocarb and to the off-peak periods. Indeed, the starting dates of the peak periods coincide with the season where the substance is spread, which is Autumn. The emergence of this two-regime pattern, alternating high concentration values during the peak periods and low concentration values during the off-peaks, is correlated with an important increase in the prosulfocarb sales as shown in Figure B.4 in Appendix B.3.2.

5.4.2. Spatial segmentation

The second step of the analysis consists in the spatial segmentation using the graph-based clustering on the monitoring stations network. This step is strictly independent from the temporal segmentation.
During the whole observation period, 420 monitoring stations only produced at least one measure. The spatial clustering algorithm was applied with a number of potential clusters varying

between 7 and 35. The minimum number of clusters is equal to the number of connected components in the graph composed of more than one station plus one cluster. There are 6 components with more than one station and we add a supplementary cluster at the initialisation of the clustering procedure. The optimal number of clusters was selected using the elbow method applied to the inertia curve. According to this criterion, illustrated in Figure in Appendix B.3.3, the best solution is made of a 15-clusters configuration. The spatial segmentation is illustrated in Figure 5.6. The algorithm based on dynamic programming 6 was used to create the partition of the station graph. To check the relevance of the homogeneity assumption formulated in section 5.2.1, let us focus on a specific temporal segment. An off-peak period, spanning between February 8, 2017 and October 4, 2017 was selected. This period was identified as a homogeneous temporal segment by the change-point detection procedure. This period is highlighted by the black rectangle in Figure 5.5. We proceeded in two steps. First we pooled all samples made during that specific period of time. From all those samples, we can identified the active clusters during that period of time, there were 13 out of 15. Then, for all active clusters, we computed the within average empirical pairwise Wasserstein distance of the active stations of a cluster and observe that for 10 clusters out of 13, this indicator is less than 0.0015, whereas the global average pairwise Wasserstein distance for the 149 stations is 0.003. This suggests that the distance chosen for our station graph is indeed a good proxy of the homogeneity in the concentration space. Additional comments can be made when we look at the geography of the region. Some clusters are overlapping with hydro-ecoregions. Hydro-ecoregions are geographic entities in which hydrographic ecosystems share common characteristics. The criteria defining them combine properties of geology, terrain and climate Wasson et al. (2002). The borders of those regions are drawn in grey in Figure . This ensures that the substances will have homogeneous dispersion properties on these clusters (see clusters 7). As expected the biggest component in Figure 5.3 is the most segmented. Some clusters are easy to identify, for instance clusters 12 corresponds to the Indre river. Cluster 10 is identified as the most western part of the Loire and its tributaries mainly the Vienne and the Creuse rivers. Clusters 1,7,9 and 10 are a little bit harder to identify. If one look closely at the map of the region, there is a high presence of small channels all across this part of the region.

5.4.3. Anomalous cluster identification

We now focus on locating spatial patterns during time segments identified in section 5.4.1. Peak periods in prosulfocarb use are less likely to produce rich spatial patterns since they correspond to a intensive overall use in the region. This why an off-peak period was investigated instead. In the rest of this case study, we selected the segment highlighted in black in figure 5.5. It is delimited by the dates February 8, 2017 and October 4, 2017. This introduces a context to the anomaly detection: a global non use of the substance.

Following the methodology proposed in 5.2.2, the scaling parameter λ_k of the aggregated data of each spatial cluster found in Section 5.4.2 was estimated. The statistics $\bar{I}_k$ was set to $1/\hat{\lambda}_k$. The Pareto front involving the two descriptors $\bar{W}_k$ and $\bar{I}_k$ was computed. It led to the cluster ranking displayed in Figure 5.7 using the *rPref* package Roocks (2016). We recall that the selected time segment corresponds to a period of non-use of prosulfocarb. From this it can be deduced that finding quantified measurements of the substance during this period is an

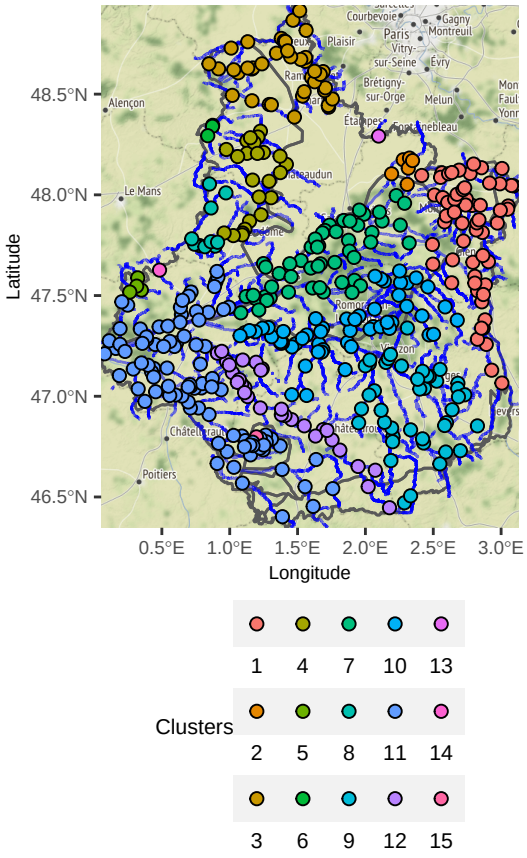

Figure 5.6: Map of geographical clusters.

anomaly. Three clusters stood out with a Pareto front levels of 1 and 2. Among them we can find on Figure 5.7:

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

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

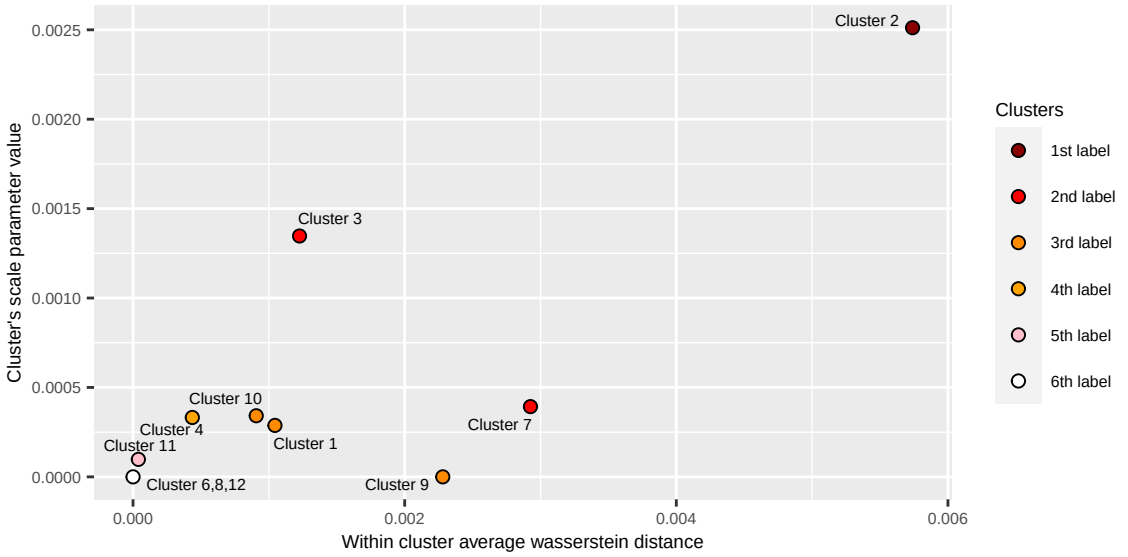

Figure 5.7: Clusters pareto front.

It is interesting to note that the Pareto front level is not uniformly distributed in the region. The three anomalous clusters are located in the north and east of the region. It could be related to the agricultural practices and land use. For the sake of the argument, we present in Appendix B.3.1 the map of barley and wheat crops in Centre-Val de Loire. In future works, we shall investigate the spatial correlation between anomalous clusters and areas with high concentration of these crops. Figure displays the Pareto front levels on the station map 5.8.

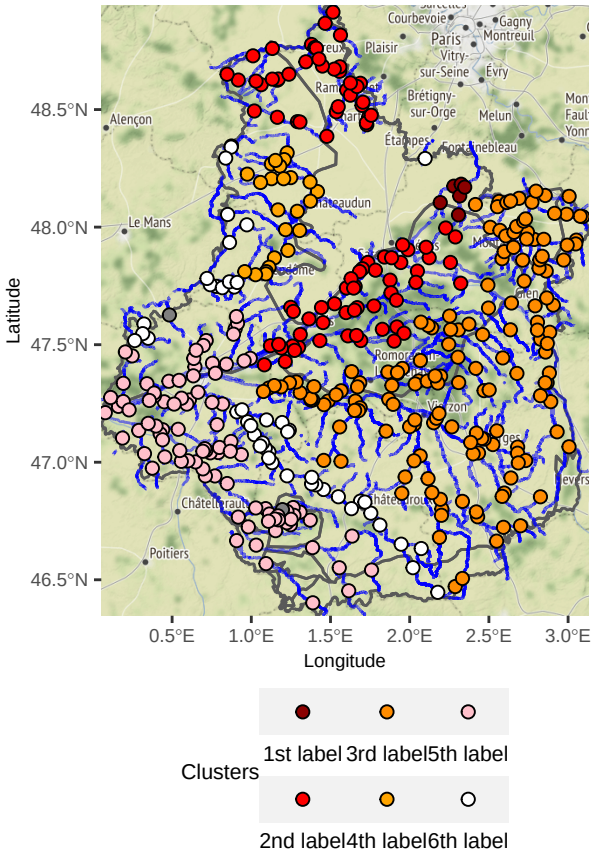

Figure 5.8: Mapped pareto front.

6. Rshiny app

Contents

All the procedure developped to detect spatio-temporal anomalies derived from Chapter 4 and 5 was implemented in an `Rshiny` application. This application is intended for the ANSES experts to assist them analyzing concentration data. Once a preprocessing is performed on the data of a substance, it can be loaded into the application. When the application starts, three tabs are available: the **Home tab** to give a overview of the susbtance informations; the **Detection tab** where the whole procedure of Chapter 5 is implemented; the **Explanatory note tab** this gives direction on how to use the application. This dociment is intended for people that don't necessary have mathematical background. It explains the events triggered by possible actions in the application.
In this Chapter, we present how the application is designed. The presentation follows the tab order in the application. In Section 6.1, we present the overview tab. We describe the elements composing the Detection tab in Section 6.2. The explanatory notice is available in C.2. This document is written in french as it is intended for the experts of the agency.

6.1. Home tab

In order to make a quick presentation of the loaded concentration dataset, we display three different elements in this tab.
The first element is text information where the following precisions are given:

- The substance's name on which the detection is performed.
- The geographical region of study.
- The dates defining the period of study.
- The total number of samples made during that time period.
- The number of active stations during that time period.
- The percentage of quantified concentration results into these samples.
- The number of days where at least a sample occured, in other words the number of daily maximum concentrations.
- The percentage of quantified daily maximum concentration results.

The second element (Figure 6.1) is the plot the daily maximum concentrations in the region. Displaying the signal at the front page allow the expert to observe the temporal trends and seasonnality of the dayly maximum concentrations without having any segmentation information to influence its interpretation.
The third element (Figure 6.2) is the map of the stations that were active during the period study. The rivers are also represented. The stations are colored according to the station graph component they belong to. We use the same method described in Section 5.3.2 relting on the BDTOPO database to create the links between stations.

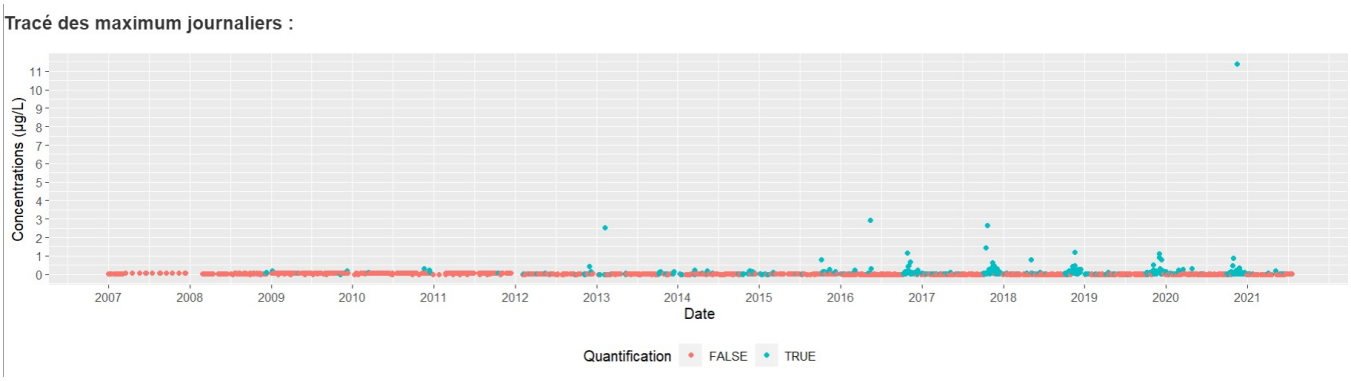

Figure 6.1: Global temporal presentation.

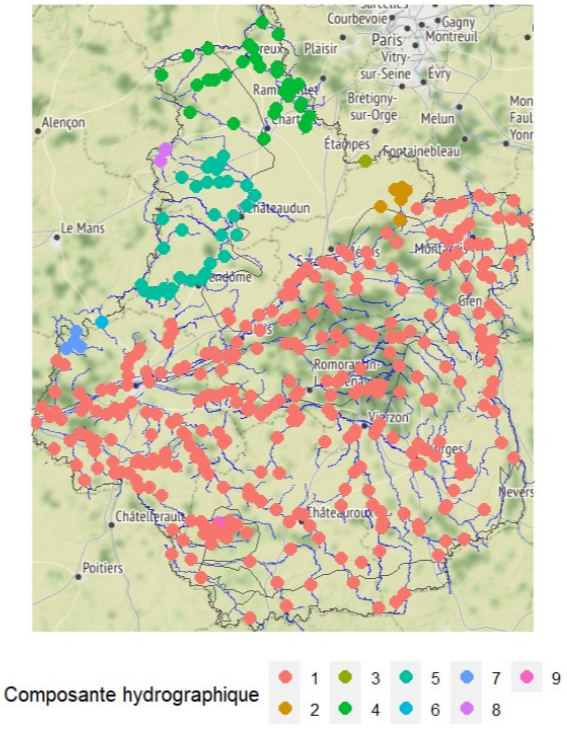

Figure 6.2: Global geographical presentation.

6.2. Detection tab

We condense the results of Chapter 5 in this single tab. The three steps of temporal detection, spatial clustering and anomaly detection are splitted in four parts.

6.2.1. Temporal detection

The temporal detection is spread across two information boxes. The first box contains the segmentation performed on the daily maximum time series. Several elements are made available to the user:

- Several segmentation results of the daily maximum time series are computed with the CROPS algorithm. For each segmentation, we have its corresponding penalty value. A curser to select the penalty value is displayed alongside with a graph of the different segmentations cost against their number of change points (see Figure 6.3). For a selected penalty value in Figure 6.3, the corresponding segmentation is highlighted in red in Figure 6.3. The red lines are the bipartite linear models obtained using the elbow method Algorithm 2. This provides the indication of what would be the optimal number of change-points. The application starts on the penalty value assoociated to the segmentation which cost is located on the elbow position. We selected a non optimal segmentation on purpose in the figures.

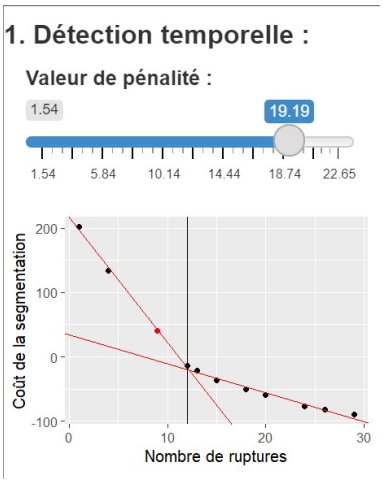

Figure 6.3: Penalty choice and corresponding segmentation information.

- Once the penalty value is set, the resulting segmentation on the daily maximum concentration signal is displayed as in Figure 6.4. This element is reactive to the penalty value and update according to the penalty value. It is also possible to select a segment in the signal. In this case, it is highlighted in black. It is the case in Figure 6.4. Lots of concentration values are under the quantification threshold. If some high concentration values occured in the signal, it can be complicated to distinguish low values of concentration. Hence, we added the possibility to plot the time series in logarithmic scale to obtain a better visualization. The concentrations in Figure 6.4 are plotted in logarithmic scale.

The second information box is dependant of the segment selected in Figure 6.4. Addtionnal informations specific to this segment are displayed:

- The following informations are given in the form of text: the dates that define the segment temporal boarders; the number of daily maximum concentration values inside the

Figure 6.4: Plot of the resulting segmentation.

segment; the quantification percentage of the segment; the number of active stations inside the temporal period defined by the segment; the minimum, the mean, the median and the maximum values of daily maximum concentrations inside the segment.

- The goodness of fit is assessed by comparing the parametric cumulative distribution function to the empirical one. The plot is illustrated in Figure 6.5, vertical black lines are drawned at the LOQ values. The empircal cdf is drawned in blue and the one obtained with the parametric model is drawned in red. Using the application confirms that the more data are present in a segment, the better the goodness of fit is. The segment selected in Figure 6.4 contained few measurments which implies the quality of the fit presented in Figure 6.5.

- The last information is a seasonnal plot. When a segment is selected, we provide a comparison of the violin plot of this segment with similar time periods on the previous and following years. For instance in Figure 6.4, the selected segment spans from the 22nd of January to the 5th of October of the year 2018. We represented the violin plot of the daily maximum concentrations from the 22nd of January to the 5th of October of each year available in Figure 6.5. 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 information box.

This two boxes encompasses the whole temporal detection procedure. The selection of the segment in Figure 6.4 determines not only the information in the second box but also all the informations that are presented in the spatial clustering and anomaly detection steps.

6.2.2. Spatial clustering and anomaly detection

Just as the temporal segmentation, the spatial clustering and the anomaly detection steps are condensed into two information boxes. The first one is a box with a map and several other elements.

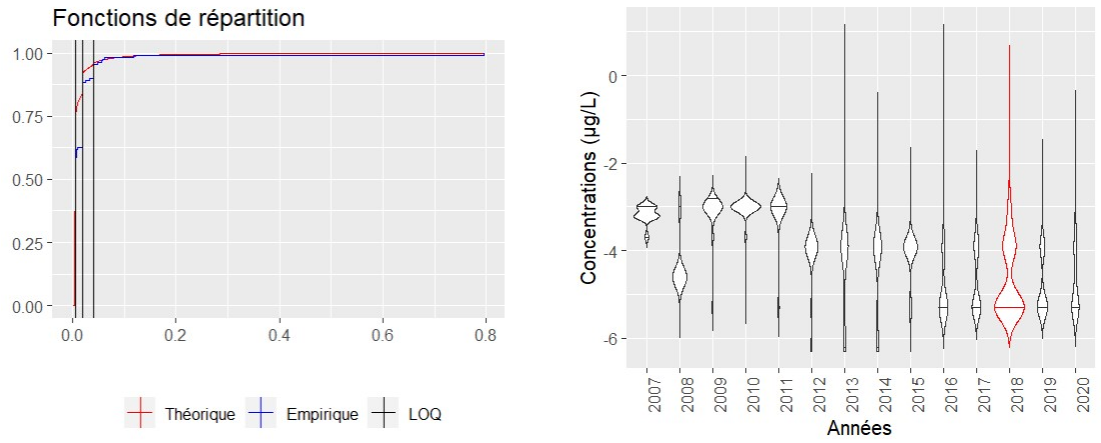

Figure 6.5: Informations on the selected segment.

A drop-down menu is present to choose the information to display on the map. As stated previously, the information is dependant of the temporal segment selected in Section 6.2.1. The user can opt for four different options: the information of activity of a station, the stations are colored differently if they made a sample or not during the segment time period; the information of the station graph components, stations are colored according to the component they belong to in the graph of stations; the cluster information, stations are colored according to which cluster they belong to (see Figure 6.6); the pareto front information, stations are colored according to their Pareto front level (see Figure 6.7). The map of the graph component is redundant with the Home tab information but it was still implemented in this box as well to avoid changing tab to get that information. The graph clustering is obtained using the clustering methods presented in Algorithm 6.
Once the information is set on either the clustering information or the Pareto front levels, it is dependant on the number of cluster information. Several clustering results were imported in the application with their respective number of clusters ranging between two values. The heuristic we use to determine the minimal and maximal values of clusters in the clustering is presented in Appendix C.1. This number can be selected with a curser located on the left in Figure 6.6 and 6.7. The map update automatically with the number of clusters information.
Additionally, there is a cluster selection feature when selecting on a station. The cluster in which the station is located is highlighted in red as in Figure 6.7. The last box of the tab is dependnanf of this choice. The last box is composed of three windows:

- The station concentration window: once a station has been selected, this window displays the samples it made during the time period of the selected segment. An illustration is provided in Figure 6.8
- The clusters Pareto plot window: clusters are plotted according to their criterion values defined in Section 5.2.2 (see Figure 6.9). The selected cluster in the map is also highlighted in red in the plot.
- Additionnal text informations are provided on all samples made by the stations belonging to the selected cluster in the last window such as: the total number of samples made in

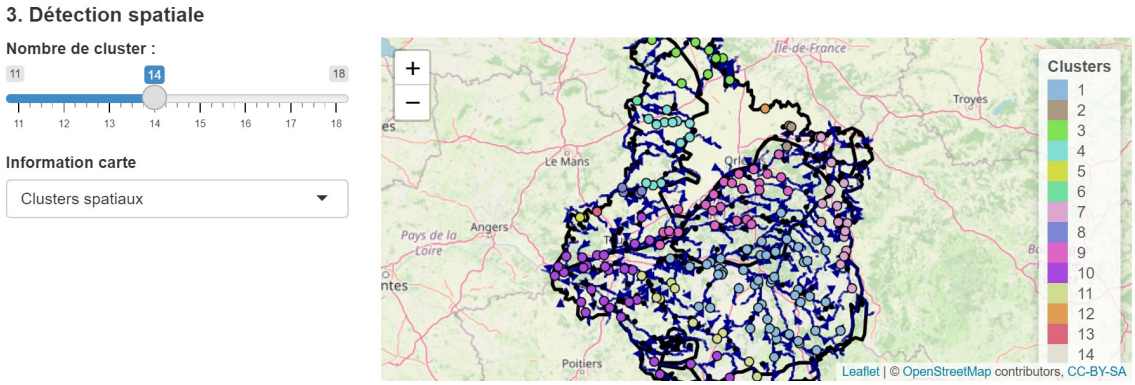

Figure 6.6: Map displaying the clusters. The clustering selected is composed of 14 clusters.

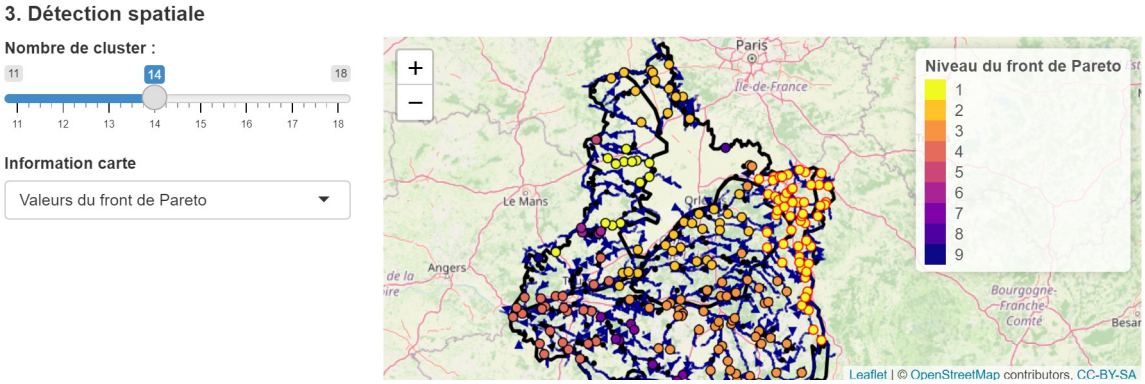

Figure 6.7: Map displaying the Pareto front values of each cluster.

that cluster in the time period selected; the percentage of quantification of these samples; the number of stations composing the cluster; the minimum, the mean, the median and the maximum of concentration values in the cluster; the LOQ values present in that cluster with the information of the most frequent LOQ; the ID of the station that has the highest quantification rate with its associated percentage of quantification rate and the number of samples that were performed. The LOQ values of a cluster provide an overview of the equipment quality that is installed on the stations.

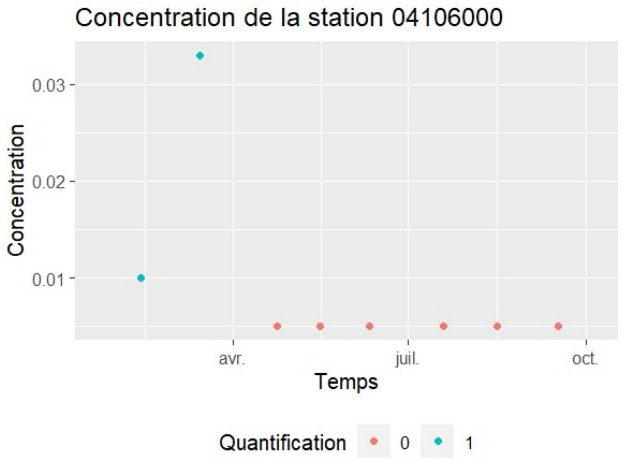

Figure 6.8: Selected station sample values during the selected temporal segment.

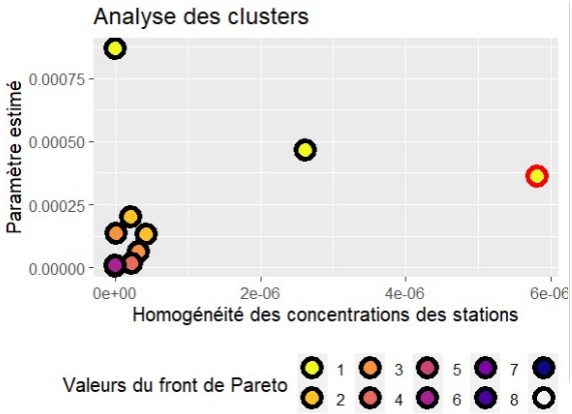

Figure 6.9: Plot of the Pareto front.

7. Conclusion

This thesis provide a procedure to extract spatio temporal informations from environmental data. We saw in Chapter 2 all the challenges that the ANSES is faced with. The mix of sampling heterogeneity and spatio-temporal heterogeneity combined with specific characteristic of censorship for concentration data demands the construction of sophisticated modeling and organized implementation. Chapters 3, 4 and 5 cover full description of our exploratory process. Our procedure is driven by temporal detection. We aim at identifying homogenous periods of time in the observed concentrations found in the environment. Chapter 3 reviewed the state of art of change-point detection methods that are used to find the homogeneous segments. In Chapter 4, we focused on studying the effects of censorship on change-point detection methods and we derived a specific optimiation method that seemed suited for our modeling. Chapter 5 provide a practical implementation of our spatio-temporal detection method. From the change-point detection constructed in Chapter 4, we can deduce the farming activities from the resulting temporal regimes using aggregated information under the form of daily maximum concentration. Once we select a specific time regime, we use the environment structure to derive spatial information. The stations monitoring surface water quality were modeled according to a graph. The links of this graph are determined by the spatial information of the riveer system of the geographical area of study. Stations were clustered according to the graph structure. The resulting clusters constitute the aggregated spatial information. A comparison of these clusters is made to identify the most anomalous ones. Chapter 6 is the practical implementation of this procedure into a `Rshiny` application. The design of this application is the results of discussion with the experts of the ANSES.

We saw that it is possible to extract information from data whose properties make modelling difficult. It seems that some natural future developments emerge from this work. The first axis of investigation regroups all works that aim to improve the modeliing of concentration data. We can push pesticids analysis forward by introducing a multivariate modeling taking simultaneous substances into account. Such methods are presented in Chapter 2 and introduced in Pickering (2016). This broadens the scope of monitoring to substances associations. The comparison of 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 use. The evolution of the spatial distribution of anomalous clusters in time can also be another crucial point. We showed in Chapter 5 that the spatial distribution of Pareto levels didn't seem to be uniform. Analysing the time series of the clusters' Pareto levels could uncover additional informations.

Although many models and methods improvments can be done, it doesn't tackle the issue of sampling. The second axis of work consists in building another sampling procedure. This procedure has to dampen the spatio-temporal heterogeneity in the collect of concentration data. This irregular sampling prevents from observing the dynamics of dispersion of a substance in space and time. The question of stations lcoations and sampling rythms can be seen as an optimal design problem Müller et al. (2011); Marsh & Ewers (2012). It can be noted that

this could prove to be a harsh task because those methods are highly dependent on the spatial structure underlying the positions of the stations. For instance, the spatial structures uncerlying stations monitoring air or surface waters quality is drastically different. The first case allows for almost any position in a given area, the latter is more constrained: stations has to be on riverside (at least near a surface water body).

References

Agriculture & Environment Research Unit (AERU) at the University of Hertfordshire. (2021). *Prosulfocarb (Ref: SC 0574).* https://sitem.herts.ac.uk/aeru/ppdb/en/Reports/557.htm. (Retrieved: March 1, 2022. Part of Lewis et al. (2016))

Andrienko, N., Andrienko, G., & Gatalsky, P. (2003, dec). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing*, *14*(6), 503–541. doi: 10.1016/s1045-926x(03)00046-6

Ansari, M. Y., Ahmad, A., Khan, S. S., Bhushan, G., & Mainuddin. (2019, jul). Spatiotemporal clustering: a review. *Artificial Intelligence Review*, *53*(4), 2381–2423. doi: 10.1007/s10462-019-09736-1

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

Antweiler, R. C., & Taylor, H. E. (2008, may). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. summary statistics. *Environmental Science & Technology*, *42*(10), 3732–3738. doi: 10.1021/es071301c

Arias-Castro, E., Candès, E. J., & Durand, A. (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics*, *39*(1), 278 – 304. Retrieved from https://doi.org/10.1214/10-AOS839 doi: 10.1214/10-AOS839

Arlot, S., & Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research*, *10*(2).

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

Bai, J. (1994, sep). LEAST SQUARES ESTIMATION OF a SHIFT IN LINEAR PROCESSES. *Journal of Time Series Analysis*, *15*(5), 453–472. doi: 10.1111/j.1467-9892.1994.tb00204.x

Baran, N., Rosenbom, A. E., Kozel, R., & Lapworth, D. (2022, oct). Pesticides and their metabolites in european groundwater: Comparing regulations and approaches to monitoring in france, denmark, england and switzerland. *Science of The Total Environment*, *842*, 156696. doi: 10.1016/j.scitotenv.2022.156696

Bardet, J.-M., Kengne, W., & Wintenberger, O. (2012, jan). Multiple breaks detection in general causal time series using penalized quasi-likelihood. *Electronic Journal of Statistics*, *6*(none). 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*, *68*, 97-122. Retrieved from `https://doi.org/10.1051/proc/202068006` doi: 10.1051/proc/202068006

Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt change: Theory and application* (Vol. 15). prentice Hall Englewood Cliffs.

Baudry, J.-P., Maugis, C., & Michel, B. (2011, apr). Slope heuristics: overview and implementation. *Statistics and Computing*, *22*(2), 455–470. doi: 10.1007/s11222-011-9236-1

Birgé, L., & Massart, P. (2006, jul). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, *138*(1-2), 33–73. 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). doi: 10.1109/ICDE.2001.914855

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*, *16*(3), 1400–1422.

Bunce, C., Carr, J. R., Nienow, P. W., Ross, N., & Killick, R. (2018, may). Ice front change of marine-terminating outlet glaciers in northwest and southeast greenland during the 21st century. *Journal of Glaciology*, *64*(246), 523–535. doi: 10.1017/jog.2018.44

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995, sep). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. doi: 10.1137/0916069

Canales, R. A., Wilson, A. M., Pearce-Walker, J. I., Verhougstraete, M. P., & Reynolds, K. A. (2018, oct). Methods for handling left-censored data in quantitative microbial risk assessment. *Applied and Environmental Microbiology*, *84*(20), e01203–18. doi: 10.1128/aem.01203-18

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*, *540*, 200-210. Retrieved from `https://www.sciencedirect.com/science/article/pii/S004896971530259X` (5th Special Issue SCARCE: River Conservation under Multiple stressors: Integration of ecological status, pollution and hydrological variability) doi: https://doi.org/10.1016/j.scitotenv.2015.06.063

Chapter 23 nonparametric tests for trend detection. (1994). In *Time series modelling of water resources and environmental systems* (pp. 853–938). Elsevier. doi: 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. doi: 10.1007/978-0-8176-4801-5

Chen, J., Kim, S.-H., & Xie, Y. (2020). S3t: A score statistic for spatiotemporal change point detection. *Sequential Analysis*, *39*(4), 563-592. Retrieved from https://doi.org/10.1080/07474946.2020.1826796 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* (Vol. 8, pp. 127–132).

Cleynen, A., & Robin, S. (2014, sep). Comparing change-point location in independent series. *Statistics and Computing*, *26*(1-2), 263–276. 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*, *7*(4), 579–588.

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*, *9*(2), 267-296. Retrieved from http://siba-ese.unisalento.it/index.php/ejasa/article/view/14726

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.*

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

de Solla, S. R., Weseloh, D. C., Hebert, C. E., & Pekarik, C. (2010, apr). 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*, *29*(7), 1476–1483. doi: 10.1002/etc.191

Einmahl, J. H. J., & McKeague, I. W. (2003). Empirical likelihood based hypothesis testing. *Bernoulli*, *9*(2), 267–290. Retrieved 2022-09-01, from http://www.jstor.org/stable/3318940

Endres, M., Roocks, P., & Kießling, W. (2015). Scalagon: an efficient skyline algorithm for all seasons. In *International conference on database systems for advanced applications* (pp. 292–308).

European Food Safety Authority. (2010, Mar). Management of left-censored data in dietary exposure assessment of chemical substances. *EFSA Journal*, *8*(3), 1557. doi: 10.2903/j.efsa.2010.1557

Faure, C., Bardet, J.-M., Olteanu, M., & Lacaille, J. (2016). Comparison of three algorithms for parametric change-point detection. In *Esann* (pp. 2–7).

Fawcett, T. (2006, jun). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi: 10.1016/j.patrec.2005.10.010

Fearnhead, P., Maidstone, R., & Letchford, A. (2018, oct). Detecting changes in slope with an l_0 penalty. *Journal of Computational and Graphical Statistics*, *28*(2), 265–275. doi: 10.1080/10618600.2018.1512868

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*, *262*, 118612.

FOMBY, T. B., & LIN, L. (2006, jan). A CHANGE POINT ANALYSIS OF THE IMPACT OF "ENVIRONMENTAL FEDERALISM" ON AGGREGATE AIR QUALITY IN THE UNITED STATES: 1940-98. *Economic Inquiry*, *44*(1), 109–120. doi: 10.1093/ei/cbj006

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75 - 174. Retrieved from `http://www.sciencedirect.com/science/article/B6TVP-4XPYXF1-1/2/99061fac6435db4343b2374d26e64ac1` 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)*, *76*(3), 495–580. Retrieved 2022-08-31, from `http://www.jstor.org/stable/24774529`

Fryzlewicz, P. (2014, dec). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, *42*(6). doi: 10.1214/14-aos1245

Gillaizeau, F., Gal, C. L., Maudet, C., Fournier, M., & Leuillet, S. (2020, sep). Méthodes de gestion des valeurs sous des seuils de détection ou de quantification. *Revue d'Épidémiologie et de Santé Publique*, *68*, S137. doi: 10.1016/j.respe.2020.03.076

Gillespie, B. W., Chen, Q., Reichert, H., Franzblau, A., Hedgeman, E., Lepkowski, J., ... Garabrant, D. H. (2010, jul). Estimating population distributions when some data are below a limit of detection by using a reverse kaplan-meier estimator. *Epidemiology*, *21*(4), S64–S70. doi: 10.1097/ede.0b013e3181ce9f08

Harchaoui, Z., Moulines, E., & Bach, F. (2008). Kernel change-point analysis. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper/2008/file/08b255a5d42b89b0585260b6f2360bdd-Paper.pdf`

Harvell, C. D., Kim, K., Burkholder, J. M., Colwell, R. R., Epstein, P. R., Grimes, D. J., ... Vasta, G. R. (1999). Emerging marine diseases–climate links and anthropogenic factors. *Science*, *285*(5433), 1505-1510. Retrieved from `https://www.science.org/doi/abs/10.1126/science.285.5433.1505` doi: 10.1126/science.285.5433.1505

Haynes, K., Eckley, I. A., & Fearnhead, P. (2014). Efficient penalty search for multiple change-point problems. *arXiv preprint arXiv:1412.3617*.

Haynes, K., Eckley, I. A., & Fearnhead, P. (2017). Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, *26*(1), 134-143. Retrieved from `https://doi.org/10.1080/10618600.2015.1116445` doi: 10.1080/10618600.2015.1116445

Haynes, K., Fearnhead, P., & Eckley, I. A. (2016, jul). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, *27*(5), 1293–1305. doi: 10.1007/s11222-016-9687-5

He, H., & Severini, T. A. (2010, aug). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, *16*(3). 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*, *73*(7-9), 1125–1141.

Hewett, P., & Ganser, G. H. (2007). A comparison of several methods for analyzing censored data. *The Annals of occupational hygiene*, *51 7*, 611-32.

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

Höök, M., & Tang, X. (2013, jan). Depletion of fossil fuels and anthropogenic climate change—a review. *Energy Policy*, *52*, 797–809. doi: 10.1016/j.enpol.2012.10.046

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

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

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions.* Wiley-Interscience. Retrieved from `https://www.ebook.de/de/product/3597498/norman_l_johnson_samuel_kotz_n_balakrishnan_continuous_univariate_distributions.html`

Jørgensen, L. F., & Stockmarr, J. (2008, dec). Groundwater monitoring in denmark: characteristics, perspectives and comparison with other countries. *Hydrogeology Journal*, *17*(4), 827–842. 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 P. A. Bernstein, Y. E. Ioannidis, R. Ramakrishnan, & D. Papadias (Eds.), *Vldb '02: Proceedings of the 28th international conference on very large databases* (p. 311-322). San Francisco: Morgan Kaufmann. Retrieved from `https://www.sciencedirect.com/science/article/pii/B9781558608696500354` doi: https://doi.org/10.1016/B978-155860869-6/50035-4

Killick, R., Fearnhead, P., & Eckley, I. A. (2012, oct). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598. 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*, *58*(4), 859. doi: 10.3349/ymj.2017.58.4.859

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, *83*(1), 79-102. Retrieved from `https://www.sciencedirect.com/science/article/pii/S030441499900023X` doi: https://doi.org/10.1016/S0304-4149(99)00023-X

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

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*, *22*(4), 1050–1064. doi: 10.1080/10807039.2015.1133242

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

Li, W., Guo, W., Luo, X., & Li, X. (2010, dec). On sliding window based change point detection for hybrid SIP DoS attack. In *2010 IEEE asia-pacific services computing conference.* IEEE. doi: 10.1109/apscc.2010.84

Li, Y., Bao, T., Shu, X., Gao, Z., Gong, J., & Zhang, K. (2021, sep). Data-driven crack behavior anomaly identification method for concrete dams in long-term service using offline and online change point detection. *Journal of Civil Structural Health Monitoring*, *11*(5), 1449–1460. 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*, *21*(3), 411–433.

Liu, C., Chen, Y., Chen, F., Zhu, P., & Chen, L. (2022, feb). Sliding window change point detection based dynamic network model inference framework for airport ground service process. *Knowledge-Based Systems*, *238*, 107701. 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* (pp. 126–135).

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*, *156*(4), 133–162.

Maimon, O., & Rokach, L. (Eds.). (2010). *Data mining and knowledge discovery handbook.* Springer US. doi: 10.1007/978-0-387-09823-4

Majumdar, A., Gelfand, A. E., & Banerjee, S. (2005). Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference*, *130*(1), 149-166. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378375804002599 (Herman Chernoff: Eightieth Birthday Felicitation Volume) 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. doi: 10.1201/9781439878125

Marchant, C., Leiva, V., Christakos, G., & Cavieres, M. F. (2018, dec). Monitoring urban environmental pollution by bivariate control charts: New methodology and case study in santiago, chile. *Environmetrics*, *30*(5), e2551. doi: 10.1002/env.2551

Marsh, C. J., & Ewers, R. M. (2012, oct). A fractal-based sampling design for ecological surveys quantifying β-diversity. *Methods in Ecology and Evolution*, *4*(1), 63–72. doi: 10.1111/j.2041-210x.2012.00256.x

Menger, J. P., Ribeiro, A. V., Potter, B. D., & Koch, R. L. (2022, jun). Change-point analysis of lambda-cyhalothrin efficacy against soybean aphid (aphis glycines matsumura): identifying practical resistance from field efficacy trials. *Pest Management Science*, *78*(8), 3638–3643. doi: 10.1002/ps.7006

Mitra, S., & Kundu, D. (2008, jul). Analysis of left censored data from the generalized exponential distribution. *Journal of Statistical Computation and Simulation*, *78*(7), 669–679. doi: 10.1080/00949650701344158

Mori, A. S., Furukawa, T., & Sasaki, T. (2012, dec). Response diversity determines the resilience of ecosystems to environmental change. *Biological Reviews*, *88*(2), 349–364. doi: 10.1111/brv.12004

Müller, W. G., Rodríguez-Díaz, J. M., & López, M. J. R. (2011, sep). Optimal design for detecting dependencies with an application in spatial ecology. *Environmetrics*, *23*(1), 37–45. doi: 10.1002/env.1132

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

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

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*, *41*, 32-48. Retrieved from https://www.sciencedirect.com/science/article/pii/S095671351300666X 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, mar). Monitoring herbicide concentrations and loads during a flood event: A comparison of grab sampling with passive sampling. *Environmental Science and Technology*, *51*(7), 3880–3891. 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: March 1, 2022)

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

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M. F., & Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, *106*(5), 808-828. doi: 10.1109/JPROC.2018.2820126

Ozgul, A., Childs, D. Z., Oli, M. K., Armitage, K. B., Blumstein, D. T., Olson, L. E., ... Coulson, T. (2010, jul). Coupled dynamics of body mass and population growth in response to environmental change. *Nature*, *466*(7305), 482–485. doi: 10.1038/nature09210

Perron, P., et al. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*, *1*(2), 278–352.

Pettitt, A. (1980, oct). Some results on estimating a change-point using non-parametric type statistics. *Journal of Statistical Computation and Simulation*, *11*(3-4), 261–272. doi: 10.1080/00949658008810413

Pickering, B. J. (2016). *Changepoint detection for acoustic sensing signals.* Lancaster University (United Kingdom).

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*.

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*, *46*(6), 900–915.

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*, *156*(4), 180–205.

Roocks, P. (2016). Computing Pareto Frontiers and Database Preferences with the rPref Package. *The R Journal*, *8*(2), 393–404. Retrieved from https://doi.org/10.32614/RJ-2016-054 doi: 10.32614/RJ-2016-054

Ryberg, K. R., Hodgkins, G. A., & Dudley, R. W. (2020, apr). Change points in annual peak streamflows: Method comparisons and historical change points in the united states. *Journal of Hydrology*, *583*, 124307. 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*.

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*, *45*(36), 6593-6606. Retrieved from https://www.sciencedirect.com/science/article/pii/S1352231011004626 doi: https://doi.org/10.1016/j.atmosenv.2011.04.073

Schaeffer, S. E. (2007, August). Graph clustering. *Computer Science Review*, *1*(1), 27–64.

Shi, X., Beaulieu, C., Killick, R., & Lund, R. (2022). Changepoint detection: An analysis of the central england temperature series. *Journal of Climate*, 1 - 46. Retrieved from https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-21-0489.1/JCLI-D-21-0489.1.xml doi: 10.1175/JCLI-D-21-0489.1

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

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*, *30*(3), 83-98. doi: 10.1109/MSP.2012.2235192

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*, 107299. Retrieved from https://www.sciencedirect.com/science/article/pii/S0165168419303494 doi: https://doi.org/10.1016/j.sigpro.2019.107299

Villani, C. (2009). *Optimal transport: old and new* (Vol. 338). 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 ressource: Agreste. (n.d.). https://agreste.agriculture.gouv.fr/agreste-web/download/methode/S-PK%20Viticulture%202019/20191217_questionnaire_PK_Viti.pdf.

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

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

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

Web ressource: ephy catalogue. (n.d.). https://ephy.anses.fr/.

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

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

Yang, T. Y., & Kuo, L. (2001, dec). Bayesian binary segmentation procedure for a poisson process with multiple changepoints. *Journal of Computational and Graphical Statistics*, *10*(4), 772–785. doi: 10.1198/106186001317243449

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz' criterion. *Statistics and Probability Letters*, *6*(3), 181-189. Retrieved from https://www.sciencedirect.com/science/article/pii/0167715288901186 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, jun). 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*, *152*(1-4), 123–131. doi: 10.1007/s10661-008-0301-y

Zhao, X., & Chu, P.-S. (2010, mar). Bayesian changepoint analysis for extreme events (typhoons, heavy rainfall, and heat waves): An RJMCMC approach. *Journal of Climate*, *23*(5), 1034–1046. doi: 10.1175/2009jcli2597.1

Zheng, K., Tan, L., Sun, Y., Wu, Y., Duan, Z., Xu, Y., & Gao, C. (2021, jul). Impacts of climate change and anthropogenic activities on vegetation change: Evidence from typical areas in china. *Ecological Indicators*, *126*, 107648. doi: 10.1016/j.ecolind.2021.107648

Zou, C., Yin, G., Feng, L., & Wang, Z. (2014, jun). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, *42*(3). doi: 10.1214/14-aos1210

Appendices

A. Chapter 4 supplementary material

This section discusses if the chosen estimators presented in Chapter 4.1 have satisfying proprieties in presence of left censored data. We will first present the hypothesis we mace on the change-point model and use some elements of the demonstration made in Lavielle (1999) to proove that the estimators $\widehat{K}$ and $\widehat{\mathcal{T}}$ still converge in the censored setting. Then, we study the convergence of a segment parameters and the importance of the initialization value in the Newton-Raphson method. In the third section, we will check if the necessary conditions to use the PELT algorithm are verified. The last part will be dedicated to the experiments on the estimation procedure proposed in 4.3.

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

The hypothesis on the change-point model are the following:

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: The maximum number of regimes may be written as $K_{\max} \geq \frac{n}{\Delta^{*}_{\boldsymbol{\tau}}}$.

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 classical hypothesis that one can find 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. We have seen in Section 4.2 that censorship could cause some problem with the identifiability. However, the introduction of a maximum value for parameter θ solved this problem. Hypothesis **H2** checks that each regime 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. The penalty value is not set in our procedure since we explore a range of values $[\beta_{min}, \beta_{max}]$ with the CROPS algorithm. However, this range of values is chosen using the BIC criterion (Yao, 1988) which verifies assumption **H4**. In change point detection, the BIC penalty can be written as:

$$\beta_{BIC} = \frac{D \log(n)}{2},$$

with D the dimension of the parameters θ and n the length of signal $\boldsymbol{y}$. The penalty range

value explored in this thesis writes under the form:

$$[\beta_{min}, \beta_{max}] = [\frac{\beta_{BIC}}{j}, \beta_{BIC} \times j],$$

with $j \in \mathbb{N}$. Hence, our penalty range is calibrated accordingly to Hypothesis **H4** because β_{min} and β_{max} also verifies its conditions.
An additionnal hypothesis, which is also the strongest one, implies that change-point locations are independent of the scale and frequency at which the data is sampled. This hypothesis will also allow us to derive the asymptotic behaviour of the estimate, when the sample size is sufficiently large. One should note here that a larger sample means a finer scale for sampling the data and not an extension of the period of observation. The scale of sampling being one of the main problem of the data we are working with, it is hard to verify. However, we can suppose that the change-points occuring in concentration data are linked with the farming activities and the phyto-pharmaceutical uses. Those events happen in fixed moments in times, thus invariant with the sampling rate.
The proof of convergence of 4.4 is based our demonstration entirely on the approach developped in Lavielle (1999). The most critical element in the proof is the condition C0(h) of Lavielle (1999). We introduce another condition. But first 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 $s > 0$,*

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

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:

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

Although $(Y_1, ..., Y_n)$ is not i.i.d., the censoring threshold makes the support of this random variables differ, the base signal is emanating from $(C_1, ..., C_n)$ that is defined in 4.1 and that is i.i.d..

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 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 :

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

where $q_{\mathcal{W}(n,n)}^{0.5}$ is the median of Weibull with parameters (n, n), $q_{\boldsymbol{x}}^{\alpha}$ is the α-th empirical quantile of the sample $\boldsymbol{x}$.
Two important points must be noted. First, all the initialisation values depend on σ. It is not problematic in our simulation tests because it is supposed known and fixed. However it stresses again the necessity of its estimation in the future (see section 4.3). Second, those estimators do not take the censorship into account. They are all biased (except if the sample $\boldsymbol{x}$ does not bear any censored values).
We tested all possible configurations with the varying values of $n = (20, 100, 500)$, $\lambda = (1/100, 1, 100)$ and a depending on a censoring rate $\alpha = (0.05, 0.25, 0.5, 0.75, 0.95)$. a was the threshold such that $\alpha\%$ of the sample was censored. The shape parameter is supposed known and fixed at $\sigma = 0.5$. For each cases, we simulated $N = 1000$ samples of left censored Weibull with 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. s

A.3. Verifying PELT assumptions

Some necessary conditions must be met before using the PELT algorithm. It can be found in Theorem 3.1 of Killick et al. (2012) and can be stated as follow:

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

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 initialisitation value: simulation results for $n = 20$.

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 inequation gives :

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

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

A.4. Convergence of $\hat{\sigma}$

The experimental protocol is the following. We simulated $N = 100$ samples of size $n = 320$ of left censored Weibull realisations. 3 change points are present in the samples at position 80, 160

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 initialisitation value: simulation results for $n = 100$.

and 240. The associated parameters for each segment are $(1, 1/100, 1/100, 1)$. Four scenarios are proposed where the σ^* and the censoring rate α varies. We test all configuration possible for $\sigma^* = (0.4, 0.8)$ and $\alpha = (25\%, 75\%)$. All the results are presented in Figures A.1 and A.2. For each of these samples, we use the estimation strategy proposed in 4.3. The penalty grid was defined by $Q = 5$ values set to $[\beta_0 = \frac{\ln n}{10}, \ldots, \beta_q, \ldots, \beta_Q = 5 \ln n]$ with the β_q being equidistant. We allowed this important range in the penalties to ensure enough distinct points for the elbow heuristic. The minimal segment size was set to 25. The choice was motivated by the computational time of the simulations, even though we have seen that the detection capacity of our method is lower with this minimal segment size. In the procedure, the first iteration $\widehat{\sigma}_0$ is computed with the `fitdistr` R package Delignette-Muller & Dutang (2015). In the first step of the estimation strategy proposed in 4.3, minimizing with respect to σ is done using Byrd et al. (1995) which allows for a box constraint for the parameter value σ (lower and upper bounds). This interval was set to $[0, 1]$. This explains that the estimated values are all inferior to 1 in Figure A.2. We defined a stopping criterion to the heuristic depending on the $\widehat{\sigma}$ values that consists in stopping when the upgrade of the new $\widehat{\sigma}$ is not superior to 10^-3.

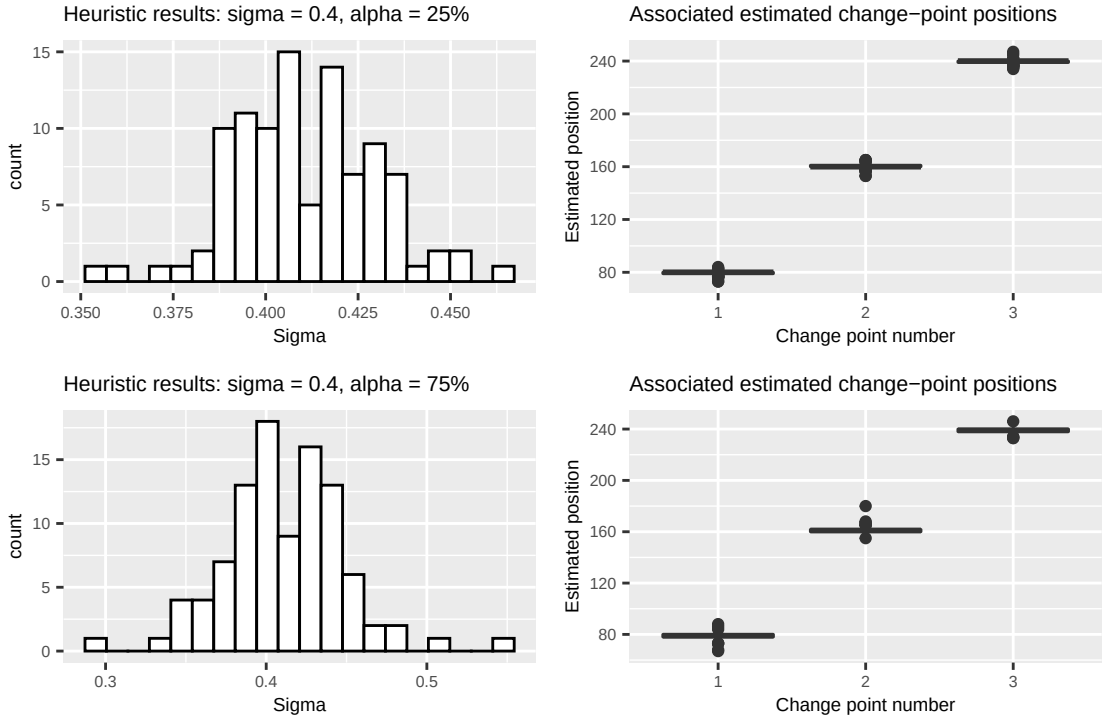

Figure A.1: Scenarios with $\sigma = 0.4$.

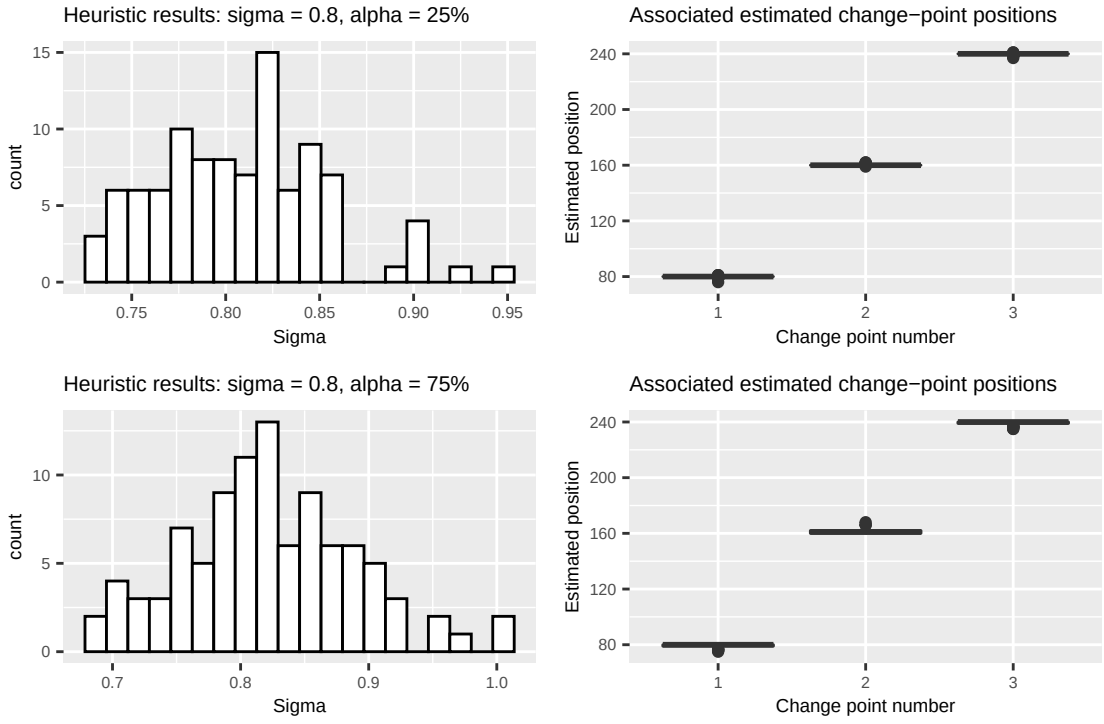

Figure A.2: Scenarios with $\sigma = 0.8$.

B. Chapter 5 supplementary material

B.1. Clustering algorithms

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

- C_m^p the m-th cluster located in component p.
- M_p the number of clusters in component $\mathcal{K}_p$.
- $Q(\mathcal{K}_p, C_m^p) = \dfrac{1}{|C_m^p|} \displaystyle\sum_{v_i, v_j \in C_m^p} d_{ij}^2$ the inertia of cluster C_m^p.
- $R_p(M_p) = \displaystyle\min_{(C_m^p)_{m=1}^{M_p}} \sum_{m=1}^{M_p} Q(\mathcal{K}_p, C_m^p)$ the best partition (in the sense of minimal inertia) of component p into M_p clusters.
- $S(l, m) = \displaystyle\min_{(M_p)_{p=1}^{l} \text{ such that } \sum_{p=1}^{l} M_p = m} \sum_{p=1}^{l} R_p(M_p)$ which is the best partition of the l first components into a total number of m clusters.

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

Algorithm 5 Clustering with greedy method:

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

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

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

Algorithm 6 Clustering by dynamic programming:

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

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

B.2. Modified empirical Wasserstein distance

The Wasserstein distance was chosen over the Kolmogorov-Smirnov or the Jensen-Shannon metric. It has the advantage of integrating in the distance calculation both the differences between the probabilities of observing different values but also the distances between those values. This is a critical point which is illustrated on a simple simulated example provided by Figure B.1. We show here three monitoring stations that have quite different behaviors. Those different behaviors are obvious both on in the temporal representation and in the histograms. However, the Kolmogorov-Smirnov distance between stations 1 and 3 is equal to the Kolmogorov-Smirnov distance between stations 1 and 2. This distance cannot capture the fact that station 2 recorded higher concentration values than station 3. On the contrary, the Wasserstein distance between stations 1 and 3 is smaller than the Wasserstein distance between stations 1 and 2.

Computing information theoretic distances/dissimilarities such as the Jensen-Shannon divergence requires estimating densities for the distributions observed at the stations. As noted earlier, few concentration records (and even fewer quantified ones) are available at the level of a station and within a time period. Therefore, density estimations based on such a small number of observations are unreliable.

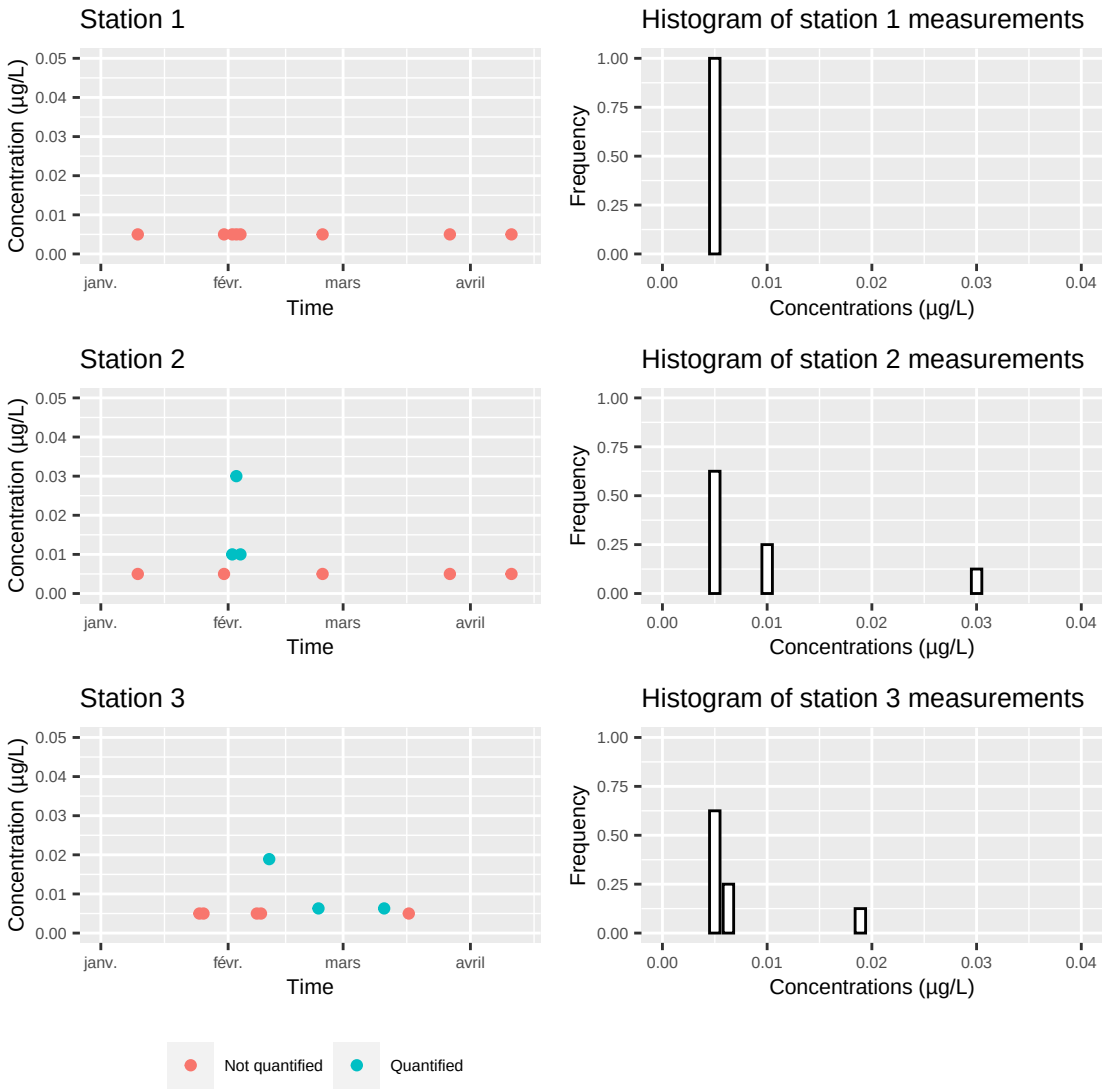

Figure B.1: Example of three stations data. The data were simulated.

The empirical 1-d Wasserstein distance used in our work is slightly adapted for left censored values. Given two samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_m)$ of sizes n and m with respective empirical c.d.f. F_n and G_m, the 1-d empirical distance writes:

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

In the case of left censored observations, the empirical c.d.f. the first non zero value is the censoring threshold. If we use the classical empirical c.d.f., it does not take into account that the potential real values of censored samples is potentially lower than this threshold. In particular, if both samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are fully censored at respective thresholds a_1 and a_2, the Wasserstein distance equals $|a_1 - a_2|$. We would like this quantity to be the smallest possible since none of the samples has any quantified values. Since the samples size for a single station

is usually very small, a reasonable assumption is to suppose that the real values under the censoring threshold are uniformly distributed. Figure B.2 illustrates the changes it implies on the empirical c.d.f.. In the previous example of $\boldsymbol{x}$ and $\boldsymbol{y}$, the adapted empirical distance gives $|a_1 - a_2|/2$.

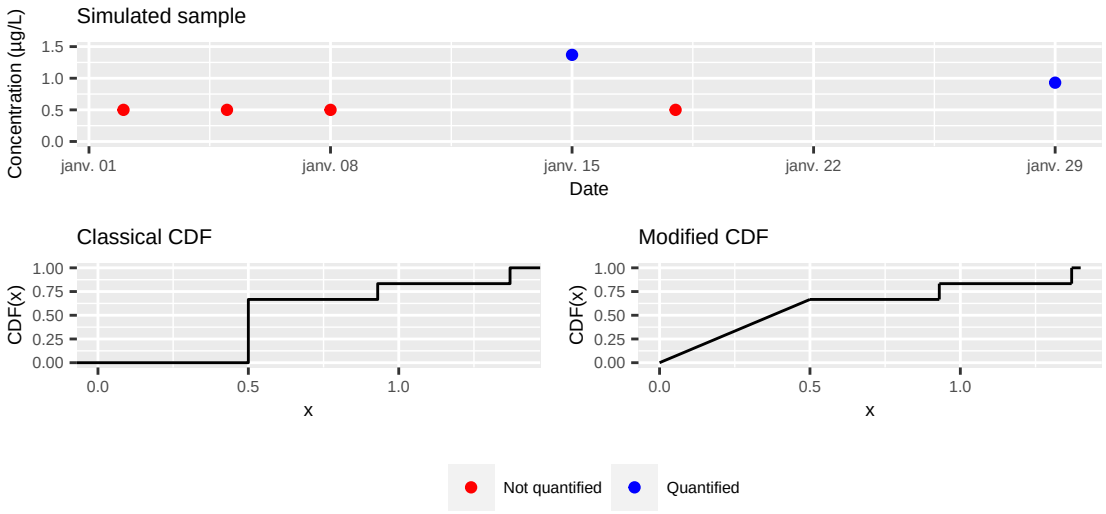

Figure B.2: Example of modified c.d.f. for the Wasserstein distance.

B.3. Supplementary Figures

B.3.1. Regional map of crops

The regional map of crops provided in Figure B.3 have been produced using data from the *registre parcellaire graphique* produced by the IGN Institut National de l'Information Géographique et Forestière (2020).

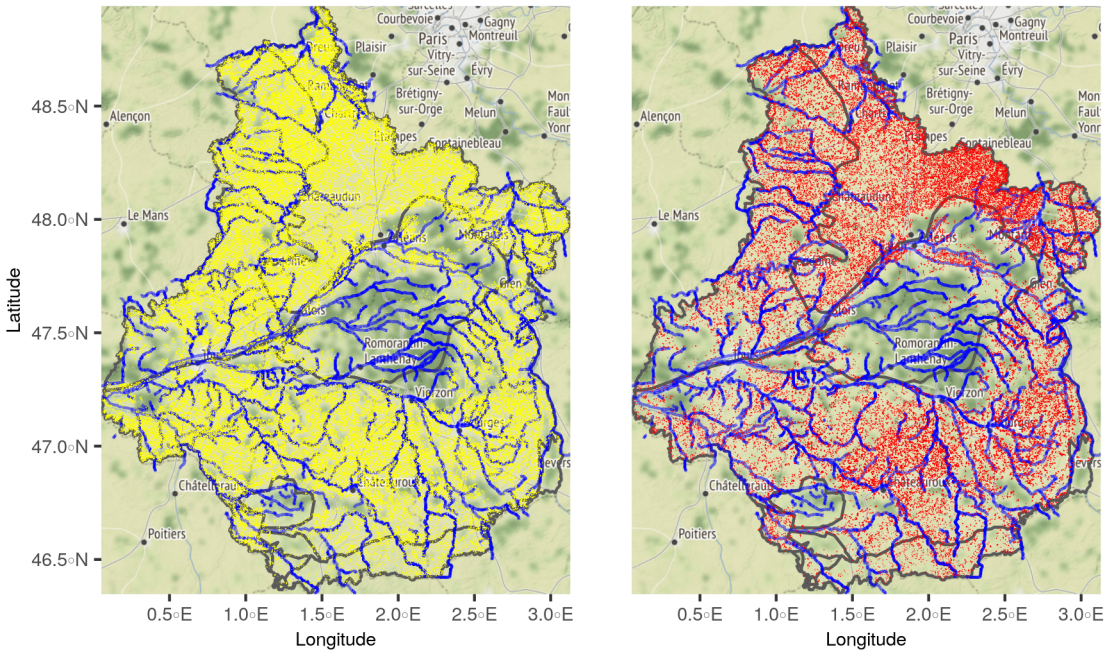

Figure B.3: Wheat (in yellow) and barley (in red) crops location in Centre-Val de Loire

B.3.2. Prosulfocarb sales

Prosulfocarb sales figures used to build Figure B.4 are made available by the *Système d'information sur l'eau* Office français de la biodiversité and Système d'Information sur l'Eau (2021).

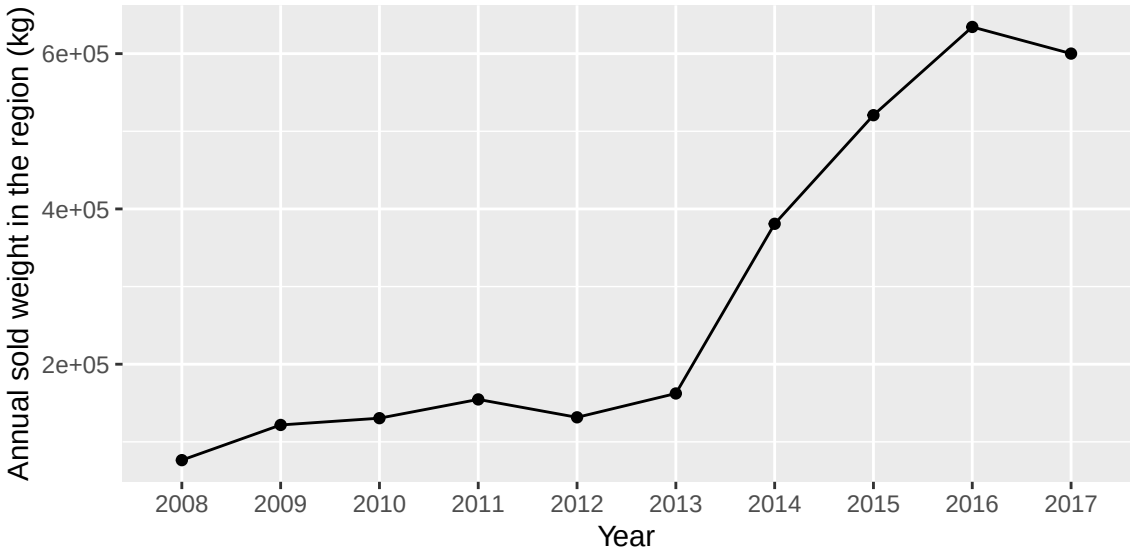

Figure B.4: Prosulfocarb sales between 2008 and 2017 in the Centre-Val de Loire region

B.3.3. All elbow methods figures

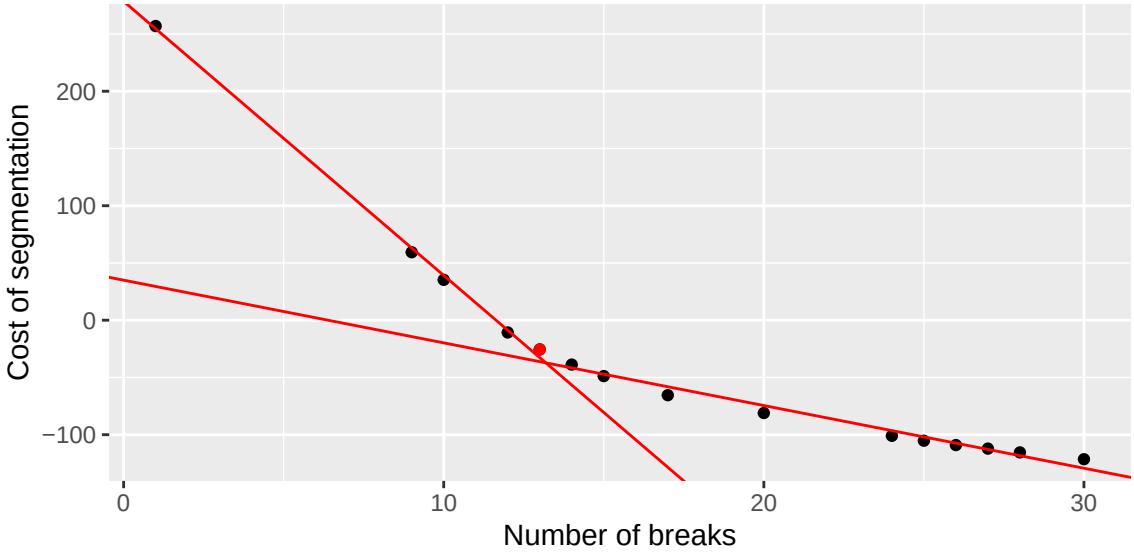

Figure B.5: Elbow method selecting the optimal segmentation of the full signal $\overline{\mathcal{D}}$.

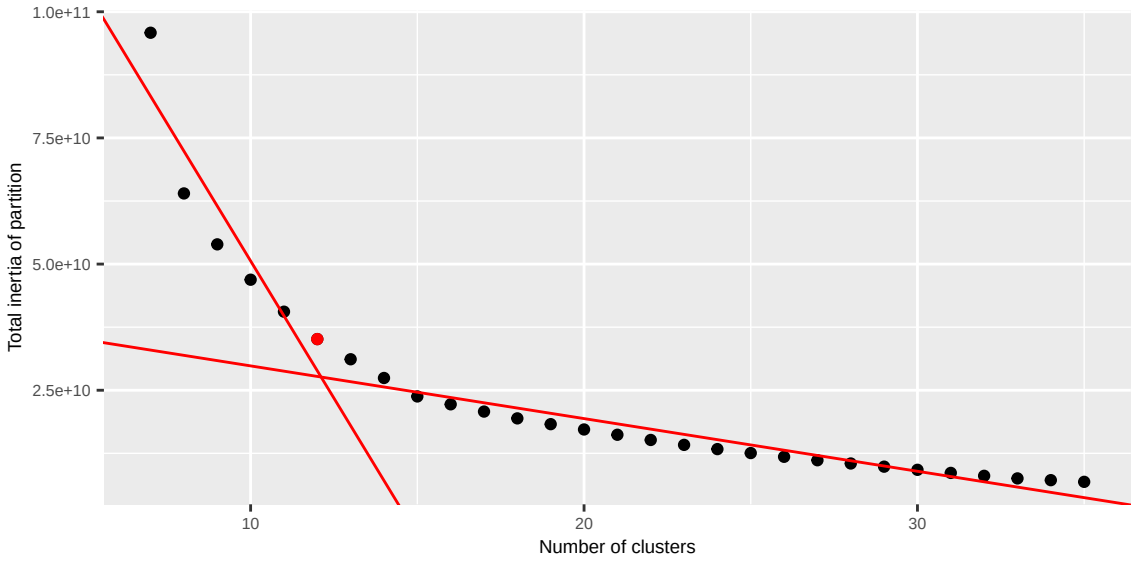

Figure B.6: Elbow method for the spatial clustering.

C. Chapter 6 supplementary material

C.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 linera behaviour. Two elbows appeared in the dicrease of the logratihm though.

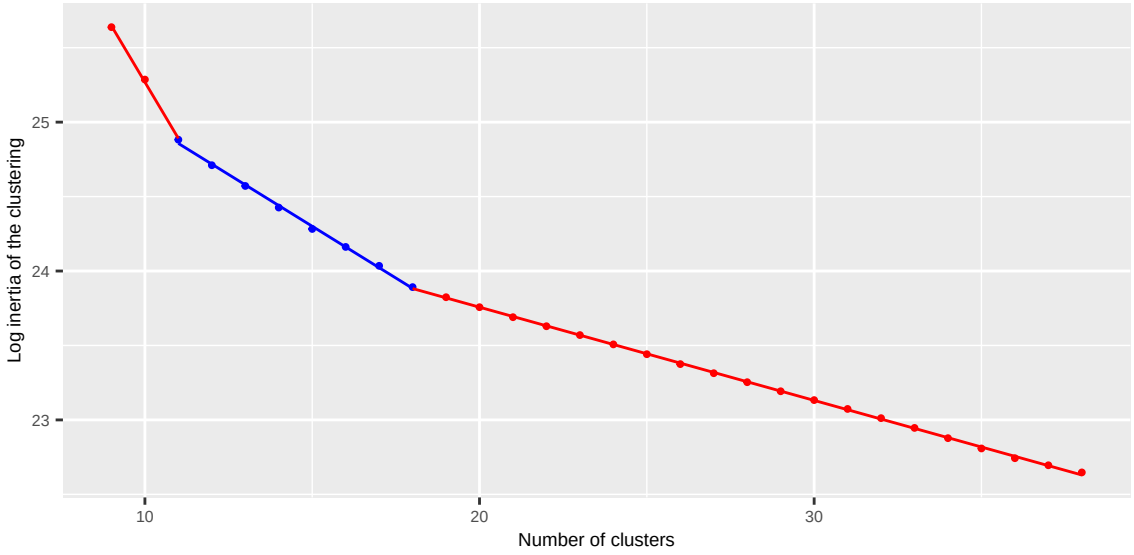

Figure C.1: Clustering candidates selected in the application.

Instead of fitting a bipartite model, we fitted the best piece-wise linear model composed of three regression models as drawned in Figure C.1. All the clustering that had their log-inertia located on the middle regression line were selected. They are coloredin blue in Figure C.1.

C.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)