POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Joint clustering of functional data from multiple sources to reconstruct historical climatic patterns in Fennoscandia

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Michele Di Sabato**

Student ID: 968811
Advisor: Prof. Simone Vantini
Co-advisors: Konrad Abramowicz
Academic Year: 2021-22

Abstract

The scientific goal of this thesis consists in the reconstruction of past changes in climate in the geographical region known as Fennoscandia during a time period of approximately 5000 years spanning from 4486 BC and AD 900. In order to achieve this goal, we use annually laminated varved sediments collected from the bottom of three lakes from Sweden (lake Kassjön) and from Finland (lakes Korttajärvi and Nautajärvi). Understanding the climatological history pertaining to the Fennoscandian Peninsula might be particularly useful to scientists, considering the lack of systematic records about the climate history of this region during past millennia. Throughout this thesis, statistical models rooted in functional data analysis and spatial statistics have been designed, implemented and compared, with the goal of clustering dependent and misaligned functional data, collected from multiple sources (i.e. the lakes). In particular, we show that it is possible to uncover a common message across the lakes, which consists in a partition of the years under study in five climate types, which are shared by all lakes and which appear gradually and coherently one after the other throughout the millennia. Lastly, we show that the shared information in the lakes consists precisely in the timing and occurrence of the five types of climate, which, on the other hand, are interpreted differently according to each lake.

Keywords: Functional clustering · Curve registration · Bagging Voronoi · Varved lake sediments · Climate change

Abstract

L'interesse scientifico di questa tesi consiste nella ricostruzione del clima nella penisola finno-scandinava in un arco di tempo di circa 5000 anni compreso tra il 4486 a.C. e il 900 d.C. facendo uso di strati annuali di sedimenti provenienti dai fondali di tre laghi in Svezia (lago Kassjön) e Finlandia (lago Korttajärvi e lago Nautajärvi). L'utilità delle analisi condotte in questo lavoro risulta particolarmente evidente considerando la mancanza di sistematiche documentazioni storiche in relazione al clima nella penisola finno-scandinava nei millenni passati. In questa tesi sono stati sviluppati, implementati e comparati modelli statistici basati su *functional data analysis* e *spatial statistics* per la classificazione non supervisionata di dati funzionali dipendenti tra di loro, disallineati e provenienti da diverse fonti. In particolare, in questa tesi mostriamo che è possibile individuare un messaggio comune ai tre laghi, che si concretizza nella suddivisione degli anni qui analizzati in cinque tipologie climatiche, valide congiuntamente per tutti i tre laghi e che si susseguono gradualmente e coerentemente tra di loro. Infine, mostriamo che ciò che accomuna i tre laghi consiste precisamente nell'occorrenza dei cambiamenti del clima, il quale, invece, si manifesta diversamente in ciascuno di essi.

Parole chiave: Clustering funzionale · Allineamento funzionale · Bagging Voronoi · Sedimenti lacustri varvati · Cambiamento climatico

Contents

Abstract **i**

Abstract **iii**

Contents **v**

Introduction **1**

1 Mathematical framework **5**

1.1 Data registration 5

1.2 Properties of (W, d) 6

1.3 Aligning multiple functions 8

1.3.1 Case of known template 8

1.3.2 Case of unknown template 8

1.4 Clustering misaligned functional data 9

1.4.1 Strictly increasing affine transformations 10

1.4.2 Square root velocity functions 13

1.4.3 Clustering misaligned functional data in the SRVF domain 16

2 The dataset **19**

2.1 The Kassjön dataset 21

2.2 The Korttajärvi and Nautajärvi data sets 24

3 Methodology **29**

3.1 First-level clustering 29

3.1.1 Initial weather cluster templates 31

3.1.2 Mitigating the over-alignment 33

3.1.3 Choosing the number of weather types k 34

3.2 Second-level clustering 43

3.2.1 Bagging Voronoi algorithm 43

3.2.2 Cell representative 45
3.2.3 Clustering model $\mathcal{M}$ 46

4 Model choices **55**
4.1 Tuning the temporal resolution parameter L 55
4.2 Selecting the number of climate types K 57
4.3 Analysis on the agreement 58
4.4 Stretching the timeline - from unsupervised to supervised 63

5 Climate interpretation **65**
5.1 Climate clusters interpretation 65
5.1.1 Climate clusters for Kassjön 67
5.1.2 Climate clusters for Korttajärvi 68
5.1.3 Climate clusters for Nautajärvi 70
5.1.4 Final remarks 71

6 Conclusions and future developments **73**

Bibliography **77**

A Appendix **81**
A.1 BVKMA 81
A.2 Evidence of agreement in the weather types 86

List of Figures **91**

List of Tables **93**

List of Symbols **96**

Acknowledgements **99**

Introduction

The recent changes in climate dating back to the mid-twentieth-century have been the cause of unfavorable outcomes and irreversible environmental events that are currently under study by the scientific community. In order to grasp the widespread effects of the recent changes in climate and to comprehend the role of human activity in this context, it is crucial to understand the climatological trends of past centuries and millennia. For future reference, we will rely on the following definitions of *weather* and *climate*: while the former is a reflection of the quickly-varying state of the atmosphere in a specific time and place, climate describes the overall prevailing weather trend over longer periods of time and in a specific location. Climate can be interpreted as the occurrence of a specific weather pattern over a long time frame: the weather of a certain year might greatly differ from the weather of the following one, however it is not uncommon for two consecutive years to share the same climate.

Due to the limited availability of global structured weather and climate monitoring before the first half of the twentieth-century, scientists have been using proxy data, such as ice-cores, tree rings and marine sediments, in an attempt to reconstruct past climate change (see Arnqvist et al. (2016)). In particular, annually laminated varved lake sediments are a promising "natural proxy" to achieve this goal, as they consist of layers of materials gradually deposited at the bottom of a specific lake. The data analyzed in this thesis have been collected from the deepest areas of three lakes, located in three different geographical regions: lake Kassjön (63° 55' N, 20° 01' E), situated in the northern Swedish region of Västerbotten, lake Korttajärvi (62° 20" N, 25° 41' E) and lake Nautajärvi (61° 48' N, 24° 41' E), both located in Finland. Each of these three sources provides one set of data, which arises from the gradual sedimentation process undergone by the lake's bed on a daily basis. The layers of sediment pertaining to each lake contain useful information about the climate around the lake itself. Each of the three sets of data is a collection of thousands of functions and each of these signals pertains to a specific year.

Given the climatological information contained in the three datasets, the goal of this

work is to investigate the presence of a common message conveyed by the three lakes and, if this message is detected, to interpret it. In particular, the scientific scope of this thesis concerns the reconstruction of past changes in climate pertaining to a broad region, encompassing all three lakes under study. From a mathematical point of view, we will achieve this goal by designing a model to cluster the years spanned by the *joint* functional dataset into a suitable number of climate types, merging the information from all lakes simultaneously. To be more precise, each year in the joint dataset will be given one specific climate label. This entails that the result of our analysis will be a single sequence of (yearly) climate labels, with which all lakes should "agree".

Evidence of a shared climate pertaining to a wider region rather than to the local surroundings of each lake has been partially uncovered in Pozzari (2017), in which functional clustering models were applied to reconstruct the local climatological history of each individual lake, thus obtaining *three* sets of yearly climate labels. In the same work, an a posteriori comparison of the sequence of climate types pertaining to each lake has been performed to prove that there exists an overlap between the local climates across the lakes. However, the author of the same work acknowledges that there are inconsistencies in the a posteriori analysis of the climates pertaining to the lakes both in the timing of the climate variations and in their interpretation. These disagreements are due to the fact that the three (local) climates have been reconstructed analyzing each lake one-at-a-time. Furthermore, lake Kassjön's dataset has been the focus of other scientific publications, such as Arnqvist et al. (2016), Abramowicz et al. (2017), Abramowicz et al. (2014). In all the cited works, each lake has been treated independently of the others, thus allowing to obtain only a *local* picture of the climatological history of the three lakes and their surrounding regions. This thesis differs from the previously cited works since the goal is to merge the information contained in all three lakes, thus obtaining a broader picture of the climate history pertaining to the Fennoscandian Peninsula.

To achieve this, multiple models have been designed, implemented and compared. In the Appendix we report an adaptation of a model proposed in Abramowicz et al. (2017) and named BVKMA (Bagging Voronoi K-Mean Alignment), since it extends the k-means algorithm to the task of clustering misaligned and temporally-dependent functional data. In particular, we designed a further extension of BVKMA to merge the information coming from all lakes simultaneously. In the Appendix we detail the issues related to BVKMA that lead us to the proposal of an alternative algorithm based on a double clustering approach, hence the name Double Clustering Bagging Voronoi (DCBV for short). The application of this model to each lake's dataset has been previously explored in Abramow-

icz et al. (2019) and Pozzari (2017). We propose an extension of DCBV to obtain a single sequence of climate labels, shared simultaneously by all lakes, in an attempt to shed light on the climate changes pertaining to a broader geographical area.

The approach used throughout this thesis is grounded in Functional Data Analysis (FDA), a branch of statistics specialized in analyzing datasets composed of functions. A common assumption in FDA is that the signals comprising a given dataset are misaligned: for this reason, suitable alignment (also known as registration) techniques have to be implemented to properly draw conclusions from the data. For this reason, a further goal of this work is to explore the literature on the topic of functional data registration and to select a suitable mathematical framework to cope with the misalignment inherent to each of the three datasets. As noted in Arnqvist et al. (2016), the rate at which sediments are gradually deposited on each lake's bed is not the same across the years. This justifies the assumption that the functions pertaining to each lake's dataset are misaligned, as explained more thoroughly in Chapter 2. An additional assumption that will be often used throughout this thesis is that there is evidence to say that there exists a temporal dependence within the datasets: indeed, the signal pertaining to a specific year is assumed to be dependent on the functions representing the neighboring ones. Note that each signal for each lake contains information regarding a specific year, hence each year is described by a triplet of functions, which can be therefore interpreted as "repeated measurements" of the climatological history of a specific year according to lake Kassjön, lake Korttajärvi and lake Nautajärvi, respecively.

This thesis is structured as follows: Chapter 1 contains the mathematical formulation pertaining to the task of registering misaligned functional data, which is the backbone of the main results proposed in this thesis. In Chapter 2 we present a more detailed description of the data collection and pre-processing steps, since the choices for the mathematical framework and the final model are deeply based on the interpretation of the functions comprising the datasets. In Chapter 3 the methodology used to reconstruct past weather and climate patterns is presented. In particular, we will propose an extension of the Double Clustering Bagging Voronoi model, previously introduced in Abramowicz et al. (2019) and applied to each of the three datasets in Pozzari (2017). Then, in Chapter 4, we suggest some suitable choices regarding the model's hyperparameters and explore some extensions of the proposed algorithm. Lastly, Chapter 5 illustrates the primary findings of this thesis, which were derived by integrating the methodologies examined in the previous chapters.

1 Mathematical framework

A considerable portion of this research has been directed towards exploring the occurrence and impacts of misalignment in the functions comprising the dataset, as well as identifying an appropriate mathematical framework to effectively address the alignment task. The objective of this chapter is to provide a comprehensive overview of the primary mathematical frameworks adopted throughout this thesis and to introduce recurring terminology and symbols. In particular, we will follow the formalism of Vantini (2012) and detail two applications of this broad framework presented in Sangalli et al. (2010) and Srivastava et al. (2011).

1.1. Data registration

Functional Data Analysis (FDA) is a branch of statistics that focuses on datasets made of curves, interpreted as realizations of random variables whose image lies in an infinite dimensional functional space (see Vantini (2012)). A functional dataset could be, for example, composed of time samples of real-valued functions on an observation interval. A common (and often necessary) step when making inference on specific phenomena using FDA, is to decouple the variability of the functions which compose the given dataset into two components: a *phase variability* and an *amplitude variability*. The former is caused by uncertainty in the measurements of the functions or even by intrinsic variability in the process which generated the observations, while the latter is the remaining source of variability once the curves have been properly aligned. It is not always necessary to remove phase variability from the dataset and it is not always the case that phase variability is even present: for example, within the functional principal component framework, it is implicitly assumed that the variability in the dataset is not due to misalignments of the time samples, but rather to other sources. When not applied in the right context, this assumption would lead to mistakenly matching evaluations of different functional data in the same abscissa and will introduce artificial inflation in variability (see Guo et al. (2022)).

The procedure of eliminating phase variability, thus allowing to focus on amplitude vari-

ability, is called curve alignment (or curve registration). According to past papers on this topic, this task can be embedded in a variety of mathematical frameworks. One of these is, for example, landmark registration: first, one finds some suitable locations in the abscissa, then the functions are aligned to match such landmarks. This approach requires manually selecting the temporal locations for each function and, in general, this could be difficult or even impossible. This is one of the motivating reasons which led to a growing number of papers aimed at constructing frameworks that automatically align functions in an unsupervised manner.

An intuitive illustration of the difference between the two sources of variability within a functional dataset is shown in Figure 1.1 inspired from Tucker et al. (2012) and obtained from data simulated using the equation $y_i(t) = z_i e^{-(t-a_i)^2/2}, t \in [-6, 6], i = 1, 2, ..., 21$, where $z_i \overset{iid}{\sim} \mathcal{N}(1, 0.05^2)$ and $a_i \overset{iid}{\sim} \mathcal{N}(0, 1.25^2)$. The variability of the functions composing this synthetic dataset is mainly due to the different locations (*phase variability*) and intensities (*amplitude variability*) of the peaks.

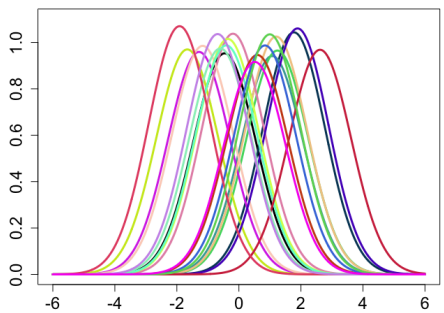


Figure 1.1: Example of phase and amplitude variability. In most practical applications the horizontal variability among the functions is much more subtle.

In more formal terms, given a metric space (F, d) where $F = \{f : \Omega \subseteq \mathbb{R} \to \Psi \subseteq \mathbb{R}^p$ with $p \geq 1\}\}$ and $d : F \times F \to \mathbb{R}_+$ is a suitable metric, aligning $f_1 \in F$ to $f_2 \in F$ means finding a function $h^* : \Omega \subseteq \mathbb{R} \to \Omega \subseteq \mathbb{R}$, called warping function, which belongs to a suitable functional space W and which is such that $f_1 \circ h^*$ is "most similar" to f_2. This optimal warping function summarizes the phase variability, leaving only the amplitude variability to be captured between $f_1 \circ h^*$ and f_2.

1.2. Properties of (W, d)

From now on we will implicitly assume that (F, d) is a suitable metric space. As many previous works on curve registration point out (e.g. Vantini (2012)), the chosen warping

space W and the metric d need to jointly satisfy a set of properties:

1. $\forall f \in F$ and $\forall h \in W$, $f \circ h \in F$;
2. W has a group structure with respect to the composition operator $\circ$, more precisely W is a subgroup of the group of the continuous automorphisms $\Omega \subseteq \mathbb{R} \to \Omega \subseteq \mathbb{R}$ (see Vantini (2012));
3. given $f_1, f_2 \in F$ and $h \in W$

$$d(f_1, f_2) = d(f_1 \circ h, f_2 \circ h) \quad (W\text{-invariance of } d) \tag{1.1}$$

Intuitively, property (1.1) ensures that applying the same warping to two different functions f_1 and f_2 does not lead them to falsely appear closer or further apart from each other.
As shown and proven in Vantini (2012), based on the metric d and the properties of the group W, assuming moreover W to be compact and that $\forall f \in F$ the map $f\circ : W \to F$ is continuous, then one can define the following pseudo-metric:

$$d_W(f_1, f_2) := \min_{h_1, h_2 \in W} d(f_1 \circ h_1, f_2 \circ h_2) \tag{1.2}$$

The space F is therefore partitioned into a quotient set, from now on $\mathcal{F}$, which can be coupled with a suitable metric [1] $d_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^+$ defined as

$$d_{\mathcal{F}}([f_1], [f_2]) := d_W(f_1, f_2) \tag{1.3}$$

It is important to point out that given $f_1, f_2 \in F$, even if there exists a pair of warping functions (h_1^*, h_2^*) such that $(h_1^*, h_2^*) = \text{argmin}_{h_1, h_2 \in W}\, d(f_1 \circ h_1, f_2 \circ h_2)$, this couple would not be unique, since $(h_1^* \circ h, h_2^* \circ h)$ with $h \in W$ would still be another minimizing couple, because of the W-invariance property (1.1). This property and the non-uniqueness of the warping functions that best align two curves f_1 and f_2 with each other, allow us without loss of generality to fix h_1 as equal to the identity warping function[2]. This means that the task of aligning two functions, denoted as f_1 and f_2, can be viewed as either fixing f_1 and aligning f_2 to f_1, or reciprocally, fixing f_2 and aligning f_1 to f_2. Following the formality of Vantini (2012), functions $\tilde{f}_1 \in [f_1]$ and $\tilde{f}_2 \in [f_2]$ are defined as *mutually-registered representatives* of equivalence classes $[f_1]$ and $[f_2]$ if and only if

[1] As proven in Vantini (2012)

[2] Thanks to the fact that $d(f_1 \circ h_1, f_2 \circ h_2) = d(f_1 \circ h_1 \circ h_1^{-1}, f_2 \circ h_2 \circ h_1^{-1}) = d(f_1, f_2 \circ h_3)$, where $h_3 := h_2 \circ h_1^{-1}$

$d(\tilde{f}_1, \tilde{f}_2) = d_{\mathcal{F}}([f_1], [f_2])$. The need to properly represent an equivalence class with one of its elements is justified by the fact that in practical applications it is more appropriate to provide results which are in the form of specific functions, rather than equivalence classes. Finally, if $\tilde{f}_1 = f_1$ and h_1 is the identity element in W, $\tilde{f}_2$ is said to be a registered version of f_2 with respect to f_1. In this case, $\tilde{f}_2$ is referred to as $\tilde{f}_{2\to 1}$ and, under the assumption of uniqueness, it can be explicitly written as

$$\tilde{f}_{2\to 1} = \underset{f\in[f_2]}{\operatorname{argmin}}\ d(f, f_1)$$

One of the conclusions drawn in Vantini (2012) is that, thanks to the partition of F into the quotient set $\mathcal{F}$, one could interpret the difference between phase and amplitude variability as the variability within and between equivalence classes, respectively.

1.3. Aligning multiple functions

Most functional datasets contain more than two functions that need to be aligned, thus a generalization of the previously explained framework needs to be tackled.

1.3.1. Case of known template

Given a functional dataset $\{f_i\}_{i=1}^n$ and a template function f_0, aligning $\{f_i\}_{i=1}^n$ with respect to f_0 simply means replacing each f_i with

$$\tilde{f}_{i\to 0} = \underset{f\in[f_i]}{\operatorname{argmin}}\ d(f, f_0)$$

or, in other words, replacing each f_i with the function in $[f_i]$ which is the closest to f_0.

1.3.2. Case of unknown template

In most practical scenarios the template f_0 is unknown, therefore Vantini (2012) proposes a comprehensive mathematical formulation for the task of aligning the given functions $\{f_i\}_{i=1}^n$, which consists in looking for a reference function $\hat{f}_0 \in F$ such that $\{\tilde{f}_i\}_{i=1}^n$ and $\{\hat{f}_0\}$ are the solution to the minimization problem:

$$\min_{\tilde{f}_i\in[f_i]\wedge\hat{f}_0\in F} \left(\sum_{i=1}^n d^2(\tilde{f}_i, \hat{f}_0) \right) \tag{1.4}$$

When it exists, the solution to (1.4) is not unique, since if $\{\tilde{f}_i\}_{i=1}^n \cup \hat{f}_0$ is a solution to (1.4), then $\{\tilde{f}_i \circ h\}_{i=1}^n \cup (\hat{f}_0 \circ h)$ is still a solution, as long as $h \in W$. Therefore, it is necessary to impose a constraint either on the template function $\hat{f}_0$ or on the warping functions $\{h_i\}_{i=1}^n$ directly. As suggested in Vantini (2012), out of all the possible equivalent representatives in $[f_0]$, one could choose $\hat{f}_0$ as the template which minimizes the average squared distance from the functions $\{f_i\}_{i=1}^n$, i.e. such that

$$\sum_{i=1}^{n} d^2(f_i, \hat{f}_0) = \min_{f \in [\tilde{f}_0]} \sum_{i=1}^{n} d^2(f_i, f) \tag{1.5}$$

In Section 1.4.1 another possible way of choosing a suitable family of warping functions will be presented.

1.4. Clustering misaligned functional data

To tackle the problem of clustering functional data, a standard approach such as the "vanilla" k-means algorithm might be inappropriate when there is evidence of misalignment among the statistical units in a functional dataset. In its essence, k-means clustering is an iterative algorithm that alternates an assignment step and a cluster template updating step, in which, respectively,

- each statistical unit is assigned a label according to which is the *closest* cluster;
- the representative of each cluster is updated as the *sample mean* of the statistical units assigned to the cluster at a specific iteration;

In the context of functional data, the choice of metric is critical and the naive pointwise sample mean might not be a viable way of updating the cluster templates, as it does not take into consideration the possible misalignment among the functions in each cluster. As long as these two steps are properly taken care of, the underlying philosophy of k-means clustering can still be applied to the functional context (see Sangalli et al. (2010); Srivastava et al. (2011); Tucker et al. (2012)), thus providing a generalization of this algorithm called k-means alignment (KMA, for future reference).
In the following subsections, we will compare two practical applications of the aforementioned mathematical framework, proposed in Sangalli et al. (2010) and Srivastava et al. (2011). The latter has been analyzed also in Tucker et al. (2012).

1.4.1. Strictly increasing affine transformations

The authors of Sangalli et al. (2010), propose to fix the set F as

$$F = \{f : \Omega \subseteq \mathbb{R} \to \Psi \subseteq \mathbb{R}^p \text{ s.t. } f \in L^2(\mathbb{R};\mathbb{R}^p),\ f' \in L^2(\mathbb{R};\mathbb{R}^p),\ f' \neq 0\} \tag{1.6}$$

and the group of warping functions W as the set of strictly increasing affine transformations

$$W = \{h : \Omega \subseteq \mathbb{R} \to \Omega \subseteq \mathbb{R} \text{ s.t. } h(t) = mt + q, \text{ with } m \in \mathbb{R}_+,\ q \in \mathbb{R}\} \tag{1.7}$$

Finally, while the theory of Vantini (2012) is based on a notion of distance, the authors of Sangalli et al. (2010) use a notion of similarity ρ defined as

$$\rho(f_1, f_2) := \frac{1}{p}\sum_{m=1}^{p} \frac{\int_{\mathbb{R}} f'_{1m}(t) f'_{2m}(t)dt}{\sqrt{\int_{\mathbb{R}} f'_{1m}(t)^2 dt}\sqrt{\int_{\mathbb{R}} f'_{2m}(t)^2 dt}} \tag{1.8}$$

where f_{1m} and f_{2m} are the m-th components of the (in general) multidimensional functions f_1 and f_2. Despite this, one can find a clear connection between the two papers by defining $d : F \times F \to \mathbb{R}_+$ as

$$d(f_1, f_2) := \sqrt{1 - \frac{1}{p}\sum_{m=1}^{p} \frac{\int_{\mathbb{R}} f'_{1m}(t) f'_{2m}(t)dt}{\sqrt{\int_{\mathbb{R}} f'_{1m}(t)^2 dt}\sqrt{\int_{\mathbb{R}} f'_{2m}(t)^2 dt}}} \tag{1.9}$$

It is important to highlight that (1.9) is a pseudo-metric. However, as noted in Vantini (2012), the same results which are valid for the case in which d is a metric, still hold if d is a pseudo-metric, as long as F is replaced with $\bar{F}$, which is the quotient set of F with respect to the relation $d(f_1, f_2) = 0$.
To summarize the main results of Sangalli et al. (2010), given a set of k template functions $\underline{\boldsymbol{\varphi}} = \{\varphi_1, \varphi_2, ..., \varphi_k\} \subset F$, $\forall\ \varphi_j \in \underline{\boldsymbol{\psi}}$, let us define the domain of attraction $\Delta_j(\underline{\boldsymbol{\varphi}})$ as the set of functions in F which, when aligned in the "best" way possible, are the closest to the j-th template out of all the others in $\underline{\boldsymbol{\varphi}}$. Namely:

$$\Delta_j(\underline{\boldsymbol{\varphi}}) := \{f \in F : \sup_{g \in W} \rho(\varphi_j, f \circ g) \ \geq\ \sup_{g \in W} \rho(\varphi_r, f \circ g)\ \ \forall\ r \neq j\} \qquad j = 1, 2, ..., k$$

From now on, the symbol $\lambda(\underline{\boldsymbol{\varphi}}, f)$ will be used to indicate the label assigned to function f out of k clusters represented by the k templates in $\underline{\boldsymbol{\varphi}}$, i.e.

$$\lambda(\underline{\boldsymbol{\varphi}}, f) = \min\{j : f \in \Delta_j(\underline{\boldsymbol{\varphi}})\}$$

Given a dataset $\{f_i\}_{i=1}^n$, the mathematical formulation of the problem consisting in clustering n misaligned functional data into k clusters according to the framework of Sangalli et al. (2010) is:

1. find a set of cluster templates $\underline{\boldsymbol{\varphi}} = \{\varphi_1, \varphi_2, ..., \varphi_k\} \subset F$ <u>*and*</u> a set of warping functions $\underline{\boldsymbol{h}} = \{h_1, h_2, ..., h_n\} \subset W$ such that

$$\sum_{i=1}^{n} \rho(\varphi_{\lambda(\underline{\boldsymbol{\varphi}}, f_i)}, f_i \circ h_i) \geq \sum_{i=1}^{n} \rho(\tilde{\varphi}_{\lambda(\underline{\tilde{\boldsymbol{\varphi}}}, f_i)}, f_i \circ \tilde{h}_i) \tag{1.10}$$

   for any set of k functions $\underline{\tilde{\boldsymbol{\varphi}}} = \{\tilde{\varphi}_1, \tilde{\varphi}_2, ..., \tilde{\varphi}_k\} \subset F$ and for any set of warping functions $\underline{\tilde{\boldsymbol{h}}} = \{\tilde{h}_1, \tilde{h}_2, ..., \tilde{h}_n\} \subset W$

2. $\forall i = 1, ..., n$ assign and align f_i to cluster $\lambda(\underline{\boldsymbol{\varphi}}, f_i)$ using the corresponding warping function h_i

As previously noted in Section 1.3.2, this problem is not well posed, since it is guaranteed that if $(\underline{\boldsymbol{\varphi}}, \underline{\boldsymbol{h}})$ is a solution of (1.10) then $(\{\varphi_1 \circ g_1, \varphi_2 \circ g_2, ..., \varphi_n \circ g_n\}, \{h_{\lambda(\underline{\boldsymbol{\varphi}}, f_n)} \circ g_1, h_{\lambda(\underline{\boldsymbol{\varphi}}, f_n)} \circ g_2, ..., h_{\lambda(\underline{\boldsymbol{\varphi}}, f_n)} \circ g_n\})$ $\forall \{g_1, ..., g_n\} \subset W$ is a solution as well, thanks to the W-invariance property. Since problem (1.10) is not analytically solvable, the authors of Sangalli et al. (2010) propose an alternative iterative procedure.
Let us define $\underline{\boldsymbol{\varphi}}_{[q-1]} = \{\varphi_{1^{[q-1]}}, \varphi_{2^{[q-1]}}, ..., \varphi_{n^{[q-1]}}\}$ as the set of cluster templates at iteration $q-1$. Moreover, fix $\{\tilde{f}_{i \to \lambda(\underline{\boldsymbol{\varphi}}_{[q-1]}, f_i)}\}_{i=1}^n$ (shortly referred to as $\{\tilde{f}_i\}_{i=1}^n$ coherently with the observations of Section 1.2) as the set of functions aligned to the representative of the corresponding cluster $\varphi_{\lambda(\underline{\boldsymbol{\varphi}}_{[q-1]}, \tilde{f}_i)}$.
At iteration q the algorithm performs the following steps:

1. **Template updating step**: find the set of new cluster templates $\underline{\boldsymbol{\varphi}}_{[q]}$ using locally estimated scatterplot smoothing (LOESS; see Cleveland (1979));

2. **Assignment and alignment step**: for each $i = 1, 2, ..., n$, $\tilde{f}_i$ is clustered to $\lambda(\underline{\boldsymbol{\varphi}}_{[q]}, \tilde{f}_i)$ and aligned to $\varphi_{\lambda(\underline{\boldsymbol{\varphi}}_{[q]}, \tilde{f}_i)}$ through an optimal warping function $h_{i^{[q]}}$;

3. **Normalization step**: for each cluster $j = 1, 2, ..., k$, the average warping function $\bar{h}_{j^{[q]}}$ is computed

$$\bar{h}_{j^{[q]}} = \frac{1}{N_{j^{[q]}}} \sum_{i:\lambda(\underline{\boldsymbol{\varphi}}_{[q]}, \tilde{f}_i)=j} h_{i^{[q]}} \tag{1.11}$$

   where $N_{j^{[q]}}$ is the cardinality of the j-th cluster at the current iteration. Then, each function $\tilde{f}_i$ is warped once again using the inverse of the average warping function

of its cluster, i.e.

$$\tilde{f}_i \leftarrow \tilde{f}_i \circ \left(\bar{h}_{\lambda(\underline{\varphi}_{[q]}, \tilde{f}_i)}\right)^{-1}$$

As explained in Section 1.3.2, the normalization step is used to select one specific warping function out of the infinitely many possible solutions to the within-cluster alignment task. Contrary to the constraint (1.5), the choice for the normalization step detailed in (1.11) ensures that the average warping undergone by curves belonging to the same cluster is the identity transformation[3]. The R package `fdakma` offers an implementation of this version of the k-mean alignment.

The mathematical framework proposed in Sangalli et al. (2010) and shortly presented in this section, offers many advantages and presents some drawbacks, concerning its application within the scope of this thesis. One of the main positive aspects is that reducing the group of warping functions to the set of affine transformations with positive slope, greatly simplifies the choice for a metric that satisfies property (1.1). Moreover, if no limitations were applied to the alignment, then the registration step would "overfit", meaning that points in the abscissa which shouldn't be linked with each other, would wrongly be matched together. Over-aligning functions happens when noise is aligned, instead of signal. Some possible drawbacks of this mathematical framework are:

1. the function d is only a pseudo-metric;
2. the computational cost of each cluster's template updating step is quite cumbersome;
3. tuning some hyperparameters involved in the algorithm is complicated considering point 2.;
4. although the original curves $\{f_i\}_{i=1}^n$ are defined on the same domain, after the alignment step their domains will change;

Although the first point does not present any issues from a theoretical point of view, a possible solution could be to, for example, impose that the warping functions in W have a slope coefficient equal to one. With this choice, it is possible to use as metric a normalized version of the L^2 distance, i.e.

$$d(f_1, f_2) = \frac{\sqrt{\int_{\mathcal{D}} (f_1(t) - f_2(t))^2 dt}}{|\mathcal{D}|} \qquad \text{(assuming } p = 1\text{)}$$

where $\mathcal{D}$ is the overlap between the domains of f_1 and f_2.

A possible solution to the second issue could be to modify the template updating step,

[3] As proven in Sangalli et al. (2010), this property holds thanks to the choice of W

replacing LOESS with other suitable regression techniques. To this goal, the R package `fdakma` has been suitably modified to enable the use of alternative template updating techniques, including smoothing splines, B-splines, as well as tree-based regression models such as LightGBM and XGBoost. According to the framework presented in this section, the clustering and alignment tasks are jointly taken care of, hence choosing a specific way of updating the cluster templates will affect the number of iterations needed for the algorithm to converge. Indeed, if the cluster templates are poorly updated, then it will take more time to the model to reach a conclusive partition of the dataset.
Regardless of the proposed possible solutions to the computational issues, the fourth point presents the most challenging problem of this mathematical framework: since the result of the alignment proposed by Sangalli et al. (2010) and implemented in the R package `fdakma` is a set of functions ultimately defined on different domains, any post-processing step aimed at interpreting the clusters might be quite problematic. While it is true that the normalization step helps in keeping the drifting away of the curves and the clusters' templates under control and despite the fact that usually phase variability is relatively mild (see Vantini (2012)), there is no guarantee that the overlap between the functions' domains after their alignment will be sufficiently wide to allow for an appropriate comparison of the signals between and within the clusters.
For these reasons, we decided to explore an alternative mathematical framework, presented in Srivastava et al. (2011) and Tucker et al. (2012).

1.4.2. Square root velocity functions

The authors of Srivastava et al. (2011) and Tucker et al. (2012) choose to work with the set of absolutely continuous functions and with warpings which are orientation-preserving diffeomorphisms. From now on we will assume that the functions in the dataset are defined on the compact interval $[0, 1]$. To be precise,

$$F = \{f : [0,1] \to \mathbb{R} \text{ s.t. } f \text{ is absolutely continuous}\} \tag{1.12}$$

$$W = \{h : [0,1] \to [0,1] \text{ s.t. } h(0) = 0,\ h(1) = 1,\ h \text{ is a diffeomorphism}\} \tag{1.13}$$

It is important to stress that the group W introduced in Section 1.4.1 is not included in the group proposed in Srivastava et al. (2011) since the identity element is the only affine transformation with positive slope, defined on $[0, 1]$ and satisfying the boundary properties in (1.13).
The choice for the metric to couple with F needs to be such that property (1.1) holds, i.e. it needs to be invariant with respect to domain warping. To do so, the authors of

Srivastava et al. (2011) and Tucker et al. (2012) propose a Riemannian metric, which is based on an inner product defined on the tangent space to a point on the manifold F. In particular, the choice of the Fisher-Rao Riemannian metric is explored: the distance $d_{FR}(f_1, f_2)$ between two functions f_1 and f_2 belonging to the manifold F is the length of the geodesic path connecting f_1 and f_2 on F. This metric is explored since, according to Srivastava et al. (2011), d_{FR} satisfies the W-invariance property. Namely, given $f_1 \in F$ and $f_2 \in F$,

$$d_{FR}(f_1, f_2) = d_{FR}(f_1 \circ h, f_2 \circ h) \qquad \forall h \in W$$

Finding the geodesic path connecting two points (functions) on the manifold F is quite involved, however in Srivastava et al. (2011) an alternative and equivalent formulation of the problem is proposed. Thanks to the absolute continuity of the functions in F and to the fundamental theorem of Lebesgue integral calculus, it is possible to map each function in F to an equivalent representative through a bijective map: given a function $f \in F$, its SRVF (square-root velocity function) representative $q : [0, 1] \to \mathbb{R}$ is a continuous map defined as

$$q(t) = \begin{cases} \frac{\dot{f}(t)}{\sqrt{|\dot{f}(t)|}} & \text{if } \dot{f}(t) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.14}$$

Since f is assumed to be absolutely continuous, its SRVF representative q belongs to $S := L^2([0, 1]; \mathbb{R})$ and f can be uniquely identified through the pair $(f(0), q)$. Notice that q alone is not sufficient to represent f, since the value of the function in a given point is needed as well.

As proven in Srivastava et al. (2011),

- under the SRVF representation, the Fisher-Rao Riemannian metric becomes the L^2 distance, i.e. $d_{FR}(f_1, f_2) = ||q_1 - q_2||_{L_2}$, where q_1 and q_2 are the representatives for f_1 and f_2;
- if q is the representative for f, then $(q, h) := (q \circ h)\sqrt{\dot{h}}$ is the representative for $f \circ h$ $\forall h \in W$;
- $\forall q_1, q_2 \in S$ and $\forall h \in W$, $||(q_1, h) - (q_2, h)||_{L_2} = ||q_1 - q_2||_{L_2}$.

Table 1.1 contains the main links between the original space of functions F and the space of the SRVF representatives, under the framework of Srivastava et al. (2011) and notation of Vantini (2012).

	Original space F	**SRVF space S**
Distance d	$d(f_1, f_2) = d_{FR}(f_1, f_2)$	$d(q_1, q_2) = \|\|q_1 - q_2\|\|_{L_2}$
Pseudo-metric d_W	$d_W(f_1, f_2) = \inf_{h \in W} d_{FR}(f_1, f_2 \circ h)$	$d_W(q_1, q_2) = \inf_{h \in W} \|\|q_1 - (q_2 \circ h)\sqrt{\dot{h}}\|\|_{L_2}$
Distance (on $\mathcal{F}$ and $\mathcal{S}$)	$d_{\mathcal{F}}([f_1], [f_2]) = d_W(f_1, f_2)$	$d_{\mathcal{S}}([q_1], [q_2]) = d_W(q_1, q_2)$

Table 1.1: The first column presents the formalism of Vantini (2012) from the point of view of the space F, while the second one is from the point of view of the square-root velocity functions.

The quotient set $\mathcal{F}$ is induced by the relation $d_W(f_1, f_2) = 0$, hence the orbit $[f]$ is defined as

$$[f] = \overline{\{f \circ h \text{ with } h \in W\}}$$

On the other hand, the orbit of the SRVF representative $q \in S$ of a function $f \in F$ is

$$[q] = \overline{\{(q, h) = (q \circ h)\sqrt{\dot{h}} \text{ with } h \in W\}}$$

i.e. it is the set of the representatives of all the functions in F which are a registered version of f. According to the notation in Srivastava et al. (2011), $\mathcal{S}$ denotes the set of all these orbits.

After having introduced this formalism, the alignment procedure described in Srivastava et al. (2011) follows the same overall reasoning of Vantini (2012): given a dataset $\{f_i\}_{i=1}^n$, the first step is to construct its SRVF representative $\{q_i\}_{i=1}^n$. The second step is to compute the Frechet mean of $\{[q_1], [q_2], ..., [q_n]\}$ under the distance $d_{\mathcal{S}}$. Note that, since $d_{\mathcal{S}}$ is defined on the quotient set of S, the Frechet mean is actually an orbit of functions in S[4]. For this reason, one needs to select one specific candidate for the orbit.

As the focus of this thesis lies not just in aligning multiple functions, but rather in clustering misaligned functional data, we will now dedicate the next section to a possible solution towards this goal proposed by Srivastava et al. (2011) and Tucker et al. (2012).

[4]this is precisely the problem of the non-uniqueness of the solution to the alignment problem, as explained in the previous sections

1.4.3. Clustering misaligned functional data in the SRVF domain

A practical implementation of both the functional clustering model and the aforementioned mathematical framework to register functional data is offered by the `fdasrvf` R package, by J. Derek Tucker and Aymeric Stamm. Algorithm 1.1 presents the main steps followed by the model proposed in Srivastava et al. (2011) and applied to the task of clustering a given dataset $\{f_1, f_2, ..., f_n\}$ under the assumption of misalignment. The best warping function in step number 4 in Algorithm 1.1 is found using a dynamic programming algorithm which keeps under control the over-alignment tendencies of the model through parameter λ: the higher λ, the less alignment is allowed. As noted in Guo et al. (2022), the over-alignment tendencies of this framework represent its most significant issue, hence parameter λ needs to be tuned appropriately (more on this in Section 3.1). Furthermore, as noted in the previous sections, a normalization technique (step 14) is used to select a specific solution to the alignment problem. Finally, notice that the template updating procedure (step 20) is performed by computing the pointwise average of the SRVF representatives of the functions belonging to a specific cluster and that the algorithm stops as soon as the maximum number of iterations is reached or if the new cluster templates are close to the old ones. The implementation of this model is suited for parallel calculations with respect to the number n of functions belonging to a cluster.
More details on the clustering model can be found in Srivastava et al. (2011) and Tucker et al. (2012).

Algorithm 1.1 K-mean alignment algorithm under the SRVF framework

1: Fix number of clusters k, maximum number of iterations M and threshold ϵ
2: Initialize cluster templates $\{\varphi_1, \varphi_2, ..., \varphi_k\}$ and compute their SRVF representatives $\{\psi_1, \psi_2, ...\psi_k\}$
3: Compute the SRVF representatives of the dataset $\{q_1, q_2, ..., q_n\}$

Find the best warping:

4: **for** $j = 1, 2, ..., k$ **do**
5: **for** $i = 1, 2, ..., n$ **do**
6: Find $h^*_{ij} = \underset{h \in W}{\text{argmin}} \; ||\psi_j - (q_i, h)||_{L_2}$
7: Compute $D_{ji} = ||\psi_j - (q_i, h^*_{ij})||_{L_2}$
8: **end for**
9: **end for**

Assignment and alignment step:

10: **for** $i = 1, 2, ..., n$ **do**
11: Assign i-th function to cluster $\alpha_i = \underset{j=1,2,...,k}{\text{argmin}} \; \{D_{ji}\}$
12: Align the i-th function accordingly: $f_i \leftarrow f_i \circ h^*_{i\alpha_i}$
13: **end for**

Normalization step:

14: **for** $j = 1, 2, ..., k$ **do**
15: Compute the Frechet mean $\overline{h}_j$ of $\{h_i\}_{i:\alpha_i=j}$ under d_{FR}
16: **end for**
17: **for** $i = 1, 2, ..., n$ **do**
18: $f_i \leftarrow f_i \circ (\bar{h}_{\alpha_i})^{-1}$
19: $q_i \leftarrow (q_i, (\bar{h}_{\alpha_i})^{-1})$
20: **end for**

Update cluster templates and check convergence:

21: **for** $j = 1, 2, ..., k$ **do**
22: Compute $\psi_j^{new} = \sum_{\{i:\alpha_i=j\}} q_i / |\{i : \alpha_i = j\}|$
23: **end for**
24: it $\leftarrow$ it $+ 1$
25: Compute $\delta = \frac{1}{n} \sum_{j=1}^{k} \frac{||\psi_j - \psi_j^{new}||_{L_2}}{||\psi_j^{new}||_{L_2}}$
26: **if** it $= M$ or $\delta \leq \epsilon$ **then**
27: **stop**
28: **else**
29: Go back to step 4
30: **end if**

2 | The dataset

In Chapter 1 we provided a brief explanation of the mathematical background on which this thesis is based. Before presenting the model proposed to uncover the climatic changes pertaining to the Fennoscandian Peninsula, it is necessary to describe how the data have been collected, motivate the assumptions raised in the introduction and explain the climatological information carried by the signals of each lake. As previously noted, the data analyzed in this thesis come from three lakes located in three different geographical regions: lake Kassjön (63° 55' N, 20° 01' E), located in the northern Swedish region of Västerbotten, lake Korttajärvi (62° 20" N, 25° 41' E) and lake Nautajärvi (61° 48' N, 24° 41' E), both located in Finland.

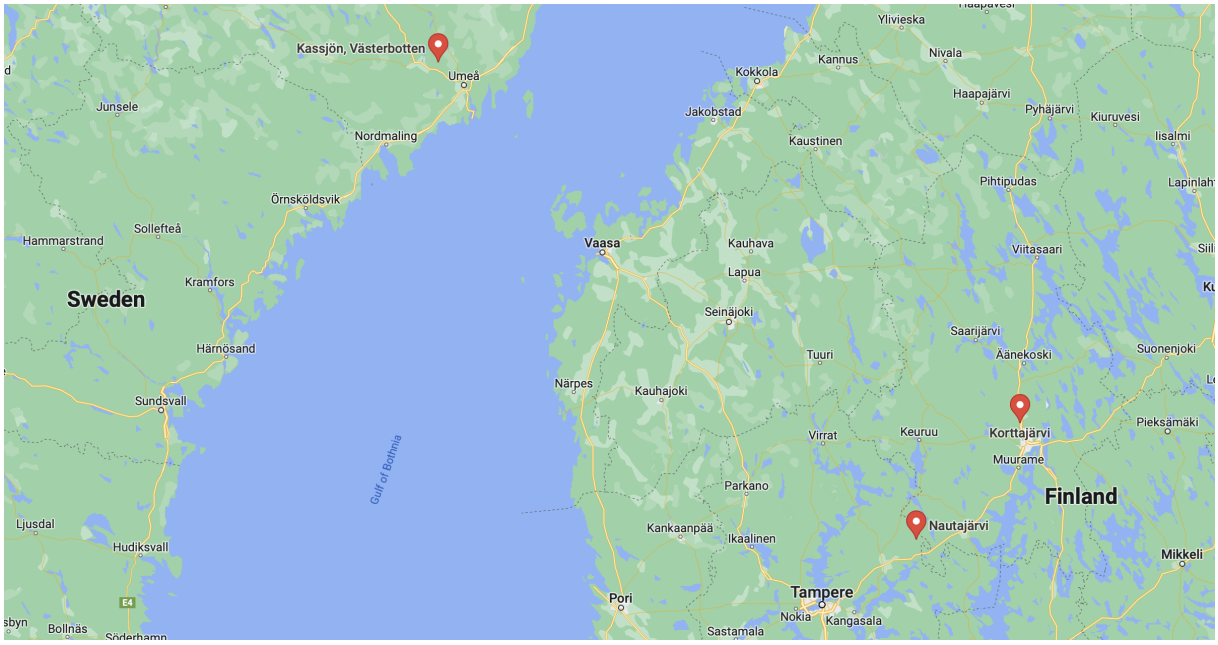
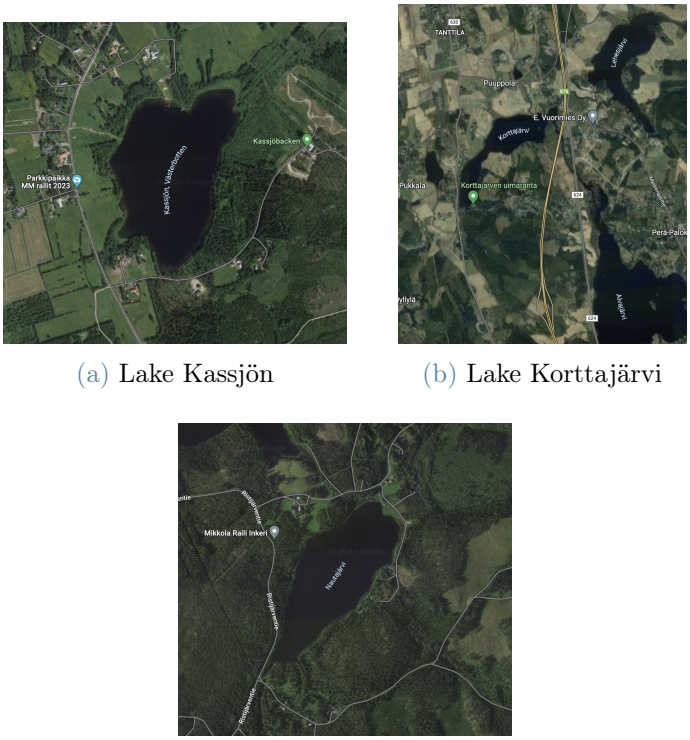


Figure 2.1: Location of lake Kassjön, Korttajärvi and Nautajärvi



(a) Lake Kassjön (b) Lake Korttajärvi

(c) Lake Nautajärvi

Figure 2.2: Satellite images of the three lakes. The presence of other lakes and water basins in the vicinity of lake Korttajärvi might have compromised its sedimentation process.

Each of the three lakes provides one set of data, which arises from the sedimentation process undergone by the lake's bed on a daily basis. The content of the sediment collected on the floor of any lake reflects its dynamic environment, its ecosystem, and its connection with the surrounding catchment. During the spring season, the accumulation of snow from the previous winter undergoes a melting process, resulting in the influx of minerogenic sediments from the surrounding regions into the lake. This natural phenomenon, known as "spring runoff", is a significant event for the sedimentation process. Throughout their life-cycle, the organisms living and dying inside the lake contribute to the sedimentation process as well, by depositing organic material. In particular, as the temperature drops and the lake starts to become ice covered during winter, the most predominant source of sedimentation will be due to such organic materials being deposited at the bottom of the lake, as the land nearby the lake itself will start to freeze. Deposit of organic matter in the lake is referred to as an autochtonous source of sedimentation, since it mainly arises from biological primary production having place inside the lake itself (as pointed out in Ojala and Alenius (2005) in the case of lake Nautajärvi). On the other hand, according to Petterson et al. (1999), mineral matter influx coming inside the lake mainly due to erosion of the surrounding catchment is an allochtonous source of sedimentation.

These are not the only ways in which minerogenic and organic materials can be deposited on the lake basin: for example, particularly heavy rains during autumn or winter might allow considerable inflows of minerogenic materials inside the lake. The presence of tributaries and nearby water basins strongly affects each lake's sedimentation process as well. Figure 2.2 shows that lake Korttajärvi is surrounded by other smaller lakes, which might have altered its settlings. As stated in Ojala and Alenius (2005), the amount of deposited sediments may vary from one year to the next according to the availability of fine-grained material in the watershed, the intensity of the spring runoff and even agricultural activity. Whatever the cause for the inflow and whichever the amount of accumulated sediments, minerogenic and organic material is slowly and constantly being deposited on top of the lake basin, thus creating varves: a varve is a collection of layers of sediments accounting for one year's worth of cumulative deposited matter. Therefore, a varve can help in reconstructing the climatological history of a specific year in terms of seasonal accumulation of various types of materials (namely minerogenic and organic).
Scientists were able to extract sequences of varves from the deepest areas of lakes Kassjön, Korttajärvi and Nautajärvi. Moreover, they manually linked each varve to a specific year. The dataset considered in this thesis is, for these reasons, a "time-series of time-series", since each year in the dataset is represented by one signal (for each lake) which is a function of time.

2.1. The Kassjön dataset

Figure 2.3 displays a portion of the dataset retrieved from lake Kassjön. Digital images of the core sediments extracted from the lake were taken and each pixel was assigned a number in the interval $[0, 255]$, to represent its grey-scaled color from black (0) to white (255). The lighter the color, the more minerogenic material is contained within a specific stratum, while the darker the shade, the higher the presence of organic material. From each varve, a sub-sample of pixels was taken and their grey-scaled values were used as representatives of the entire varve. Because of the fact that different varves have different thickness (depending on the sedimentation rate of a specific year) the size of these samples differ from varve to varve. This allows experts to take into account the sedimentation rates in their analysis.

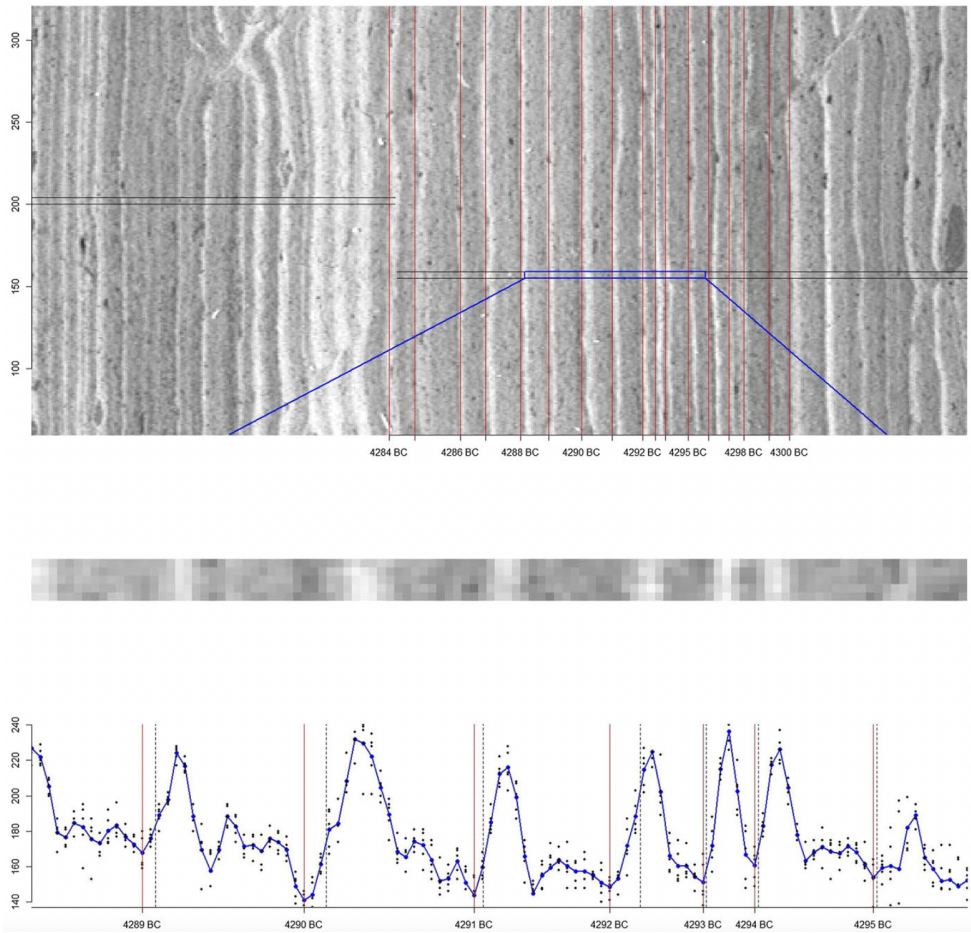


Figure 2.3: Varves' structure of lake Kassjön obtained with a digital signal processing video camera, *source:* Arnqvist et al. (2016)

The *top* image in Figure 2.3 shows sixteen varves from lake Kassjön, manually delimited through solid red vertical lines. The delimiters have been set according to the fact that during most winters the color of the varve becomes darker and darker while during spring it usually becomes much lighter. The *middle* image in Figure 2.3 shows the subsampling procedure applied to some of the sixteen varves. Finally, the *bottom* image displays the gray-scale intensities of the pixels, which comprise the raw functional dataset for lake Kassjön. The dashed black lines in the bottom figure display the original delimiters, which have then been shifted to the red lines to cope with the fact that in general winter is typically linked to the darkest shades. A varve is hence defined from the start of a spring to the end of a winter layer. This is just a formality: indeed, defining the precise layer which determines a new year is extremely difficult, therefore the easily detectable contrast between the dark color of winter layers and the bright color of spring layers can be used to define the end of a year and the start of a new one[1]. For more information on

[1]Other works (such as Tiljander et al. (2002)) define a varve from the beginning of winter to the end

the data collection from a geological perspective see Petterson et al. (1999) and Petterson et al. (1993).
Lake Kassjön provides data from 4486 BC to AD 1898 amounting for a total of 6388 varves, sixty-two of which contain missing values. Thanks to the efforts in Arnqvist et al. (2016), lake Kassjön's data have undergone some pre-processing steps, the first of which consisted in shifting the domain to the compact interval $[0, 1]$, to allow for a better comparison of differently sized varves. Secondly the data points representing each varve have been smoothed using penalized B-spline basis functions with a penalization parameter chosen through generalized cross-validation (GCV). The pre-processing of the Kassjön dataset has been performed in Arnqvist et al. (2016).
The final data after this pre-processing step are:

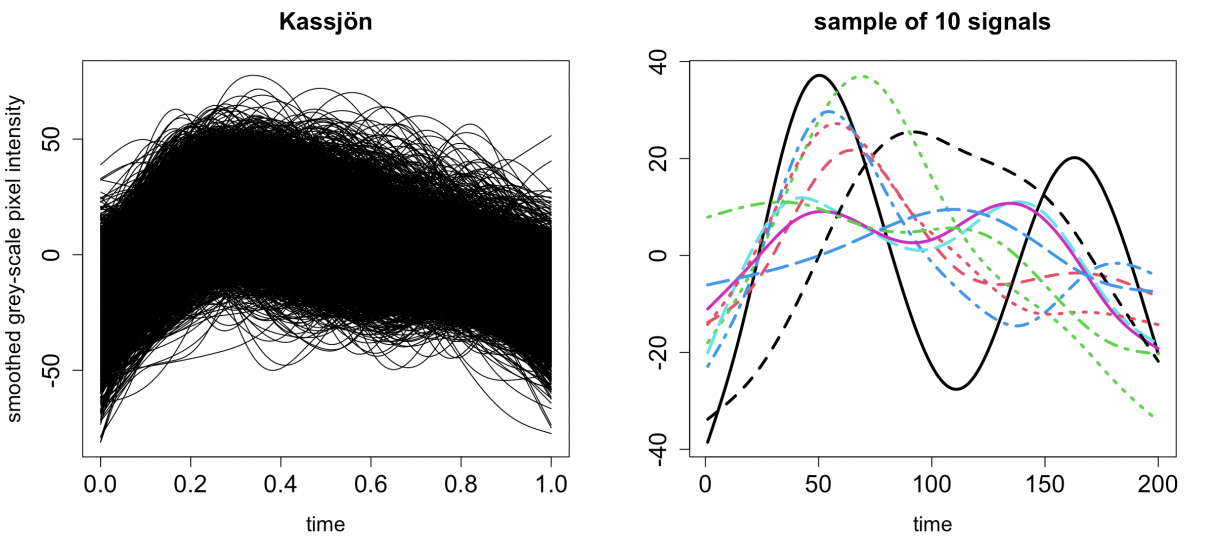


Figure 2.4: Pre-processed Kassjön data. The functions have been sampled using two hundred equally spaced points in the interval $[0, 1]$.

Each function contains information on both organic and minerogenic sedimentation on the bottom of the lake: time periods characterized by high peaks correspond to time intervals in which mainly minerogenic material was deposited (i.e. corresponding to a bright stratum in the varve as in Figure 2.4). On the contrary, periods in which the functions have a low value correspond to intervals when mostly organic material was deposited on the basin of the lake (i.e. the darker strata in Figure 2.4).
The shifting of the abscissa to $[0, 1]$ and the smoothing procedure are necessary pre-processing steps, although they will cause some loss of information in regards to the stratification rate.

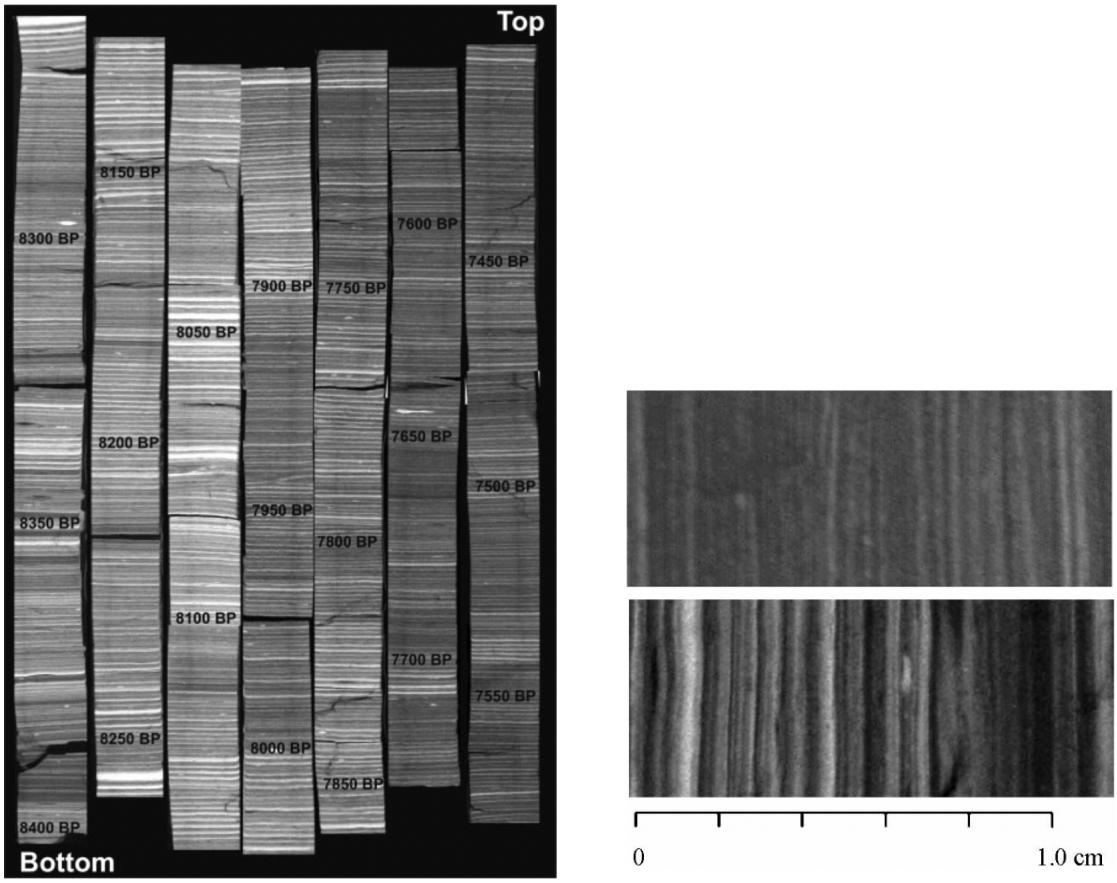
of spring.

2.2. The Korttajärvi and Nautajärvi data sets

While Kassjön's dataset is based on digital image analysis, the sequence of varves from lake Korttajärvi and Nautajärvi is based on X-ray radiography, which is a non-destructive technique instrumental in clastic-organic sediment analysis. As explained in Tiljander et al. (2002), thin slices of sediments have been positioned between an anode and a film and subjected to X-rays. Some of the energy conveyed through these rays is absorbed by the molecules of the sediments, thus creating a light-colored shadow on the film. Because of their density, minerogenic materials have a higher tendency in absorbing the X-rays (see Tiljander et al. (2002)), and hence they project a much brighter shadow on the film. Figure 2.5 shows the X-ray image of a portion of lake Nautajärvi and Korttajärvi's varves. Although these are not digital images of the varves, the interpretation of the bright and dark-colored strata is similar to the one presented in section 2.1. The raw functions which comprise lake Korttajärvi and Nautajärvi's datasets represent the gray-scaled intensity of the X-ray images of the varves. For a thorough explanation of the data gathering process using X-ray radiometry, see Tiljander et al. (2002).
The time series pertaining to lake Nautajärvi spans the time period between 7881 BC and AD 1996, for a total of 9878 years, while lake Korttajärvi provides 8393 varves, from 7580 BC to AD 900. The pre-processing of both Finnish lakes has been performed in Pozzari (2017) and involved

- smoothing the functions through penalized B-splines (GCV was used to select the regularization parameter);
- shifting the domain to the interval $[0, 1]$;
- selecting the varve delimiters such that each varve starts at spring and ends in winter;
- removing the average from the datasets.

The final data after these pre-processing steps are shown in Figure 2.6 and 2.7. The interpretation of the pre-processed functions from the Finnish lakes is the same as the one for lake Kassjön's dataset: time periods in a specific year characterized mainly by *minerogenic* inflows inside the lake, are associated with high peaks of the function describing the year. On the contrary, the lower the signal in a specific interval of a year, the higher the *organic* content of the varve's layer which describes the same interval of that particular year.



(a) lake Nautajärvi, *source:* Ojala et al. (2008)

(b) lake Korttajärvi AD 1235–1209 (*top*) and 721–706 BC (*below*), *source:* Tiljander et al. (2003)

Figure 2.5: X-ray images of varved sediments

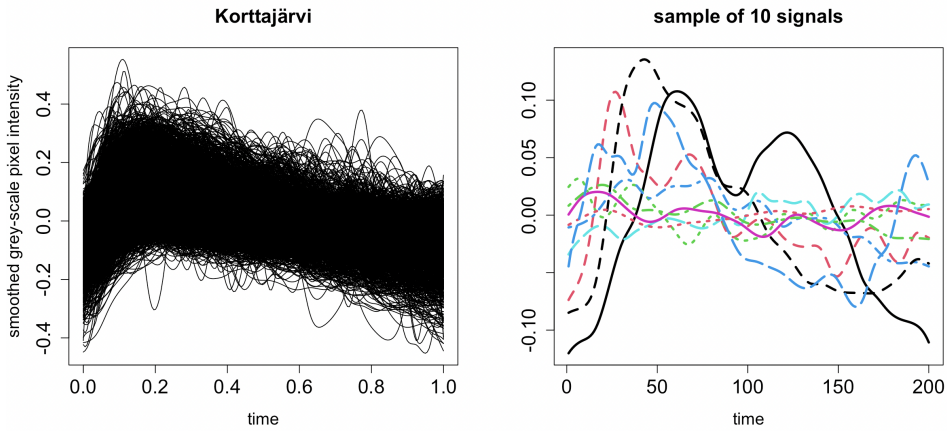


Figure 2.6: Pre-processed Korttajärvi data. The functions have been sampled using two hundred equally spaced points in the interval $[0, 1]$.

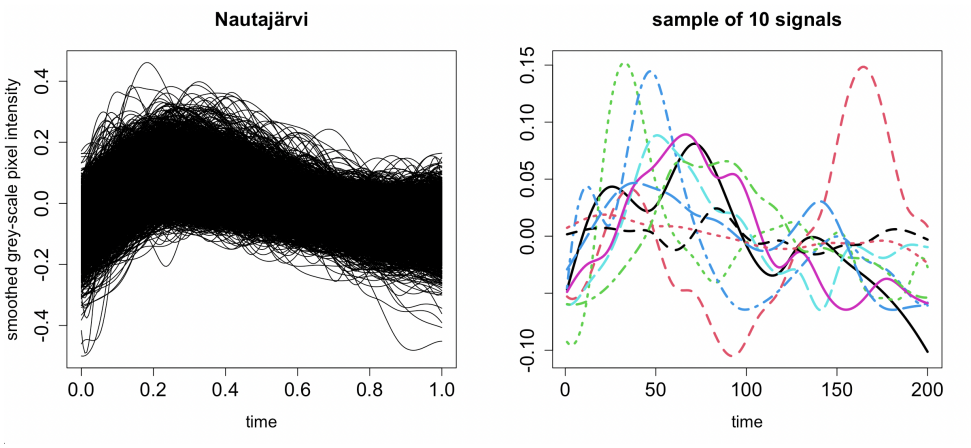


Figure 2.7: Pre-processed Nautajärvi data. The functions have been sampled using two hundred equally spaced points in the interval $[0, 1]$.

Because of the different measurement types (image analysis for the Swedish lake and X-ray radiography for the Finnish ones), the functions coming from lake Kassjön are clearly much different from the ones of the Finnish lakes. For example while the range of the signals coming from Korttajärvi and Nautajärvi are $[-0.4526, 0.4646]$ and $[-0.4907, 0.4612]$ respectively, the ones from Kassjön range in $[-81.0998, 77.5491]$. To make the three lakes more comparable, the dataset from Kassjön has been re-scaled to $[-0.4716, 0.4629]$ so that the extremes of the new range will correspond to an average of the maximum and the minimum values of Korttajärvi and Nautajärvi[2]. Moreover, a visual inspection of a subset of the curves across the lakes shows that the signals from the Finnish lakes are more noisy than the ones from Kassjön. The color and line-type of the ten signals shown on the right of Figures 2.4, 2.6 and 2.7 correspond to the same years across the different lakes. Note that the most relevant features of the signals in each dataset are the presence (or absence) and intensity of the peaks, the average value of the signal (which can be an indicator of the overall temperature for a specific year and lake) and the velocity with which a function reaches a specific peak (see Ojala and Alenius (2005)).
Finally, the time period spanned by each dataset is lake-specific: as Figure 2.8 shows, the lakes share 5271 years from 4486 BC to AD 900. Hence, from now on we will work on this time period.

[2]i.e. $-0.4716 = -(0.4526 + 0.4907)/2$ and $0.4629 = (0.4646 + 0.4612)/2$

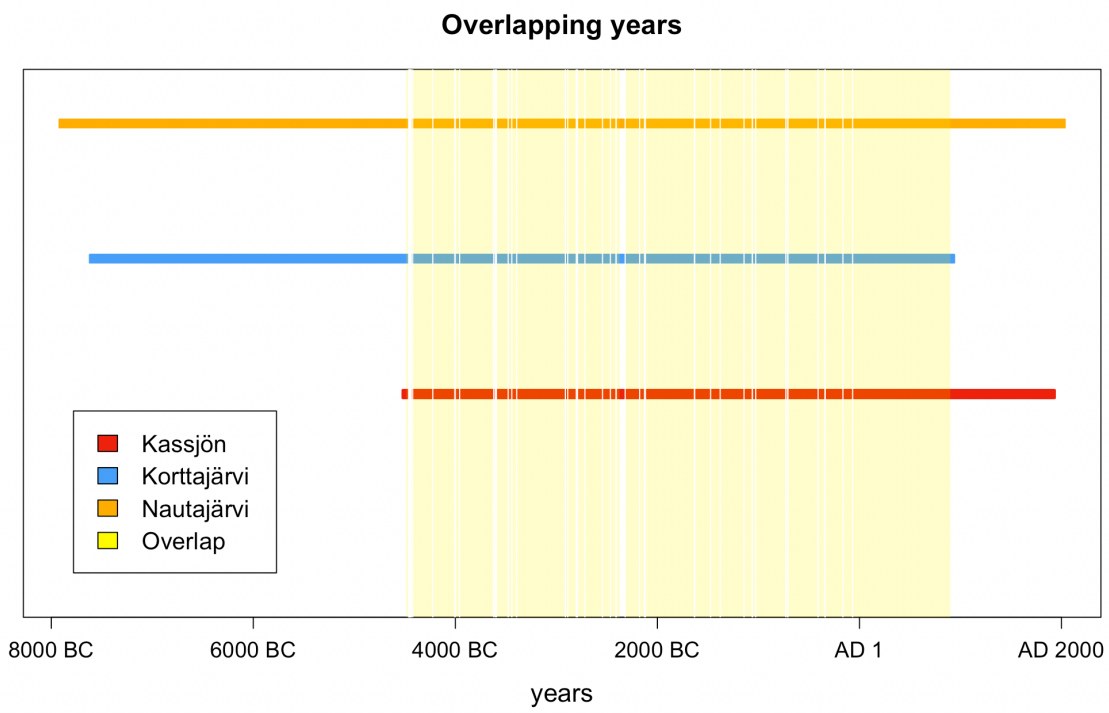


Figure 2.8: Plot of the time period spanned by each lake's data. The overlapping years are colored in yellow and the white gaps represent years which are missing data for at least one lake. Such missing years have been removed from each dataset.

3 | Methodology

In this chapter, the methodology used to uncover the climatic variations pertaining to the three lakes will be detailed. In line with previous works on the same topic (see Abramowicz et al. (2019)), a double clustering approach will be employed: the role of the first-level clustering is to label each year as belonging to a specific type of *weather*. The second-level clustering will be then used to partition the years into a chosen number of *climate* clusters, coherently with the interpretation of climate as the occurrence of a specific weather pattern over time. In short, the first-level clustering will reconstruct weather and the second will cluster climate. This model has therefore been named Double Clustering Bagging Voronoi (DCBV) and has been used both in Pozzari (2017) and Abramowicz et al. (2019). We stress that while previous works on the same topic applied DCBV *one-lake-at-a-time*, in this thesis we propose an extension of this model to the *multi-lake setting*.
Because of the geographically and temporally local nature of weather, in the first-level clustering each lake will be treated independently of the other two, while the aim of the second-level clustering will be to merge the information coming from the three geographical locations and provide a unique sequence of climate labels with which all lakes agree. Furthermore, the first-level clustering model is designed to take into consideration the misalignment among the functions comprising each of the three datasets, while the second-level clustering model will be used to account for the temporal dependence of neighboring years.
Neither the number of weather types nor the number of climate types is known a priori.

3.1. First-level clustering

From now on, $n = 5271$ will indicate the number of years in the joint dataset, the i-th function will be denoted as $\boldsymbol{f}_i(t)$ and $\{t_j\}_{j=1}^{200}$ will indicate the equispaced grid in $[0, 1]$ on which the functions $\{\boldsymbol{f}_i(t)\}_{i=1}^{n}$ have been sampled. In particular,

$$\boldsymbol{f}_i(t_j) = \begin{bmatrix} f_i^1(t_j) \\ f_i^2(t_j) \\ f_i^3(t_j) \end{bmatrix} \in \mathbb{R}^3 \qquad \forall i = 1, 2, ..., n \text{ and } \forall j = 1, 2, ..., 200$$

where $f_i^1(t_j), f_i^2(t_j), f_i^3(t_j)$ represent the functions pertaining to the i-th year, sampled at time t_j and regarding lake Kassjön, Korttajärvi and Nautajärvi respectively[1].
The sedimentation rate varies from year to year: for example, if one year is characterized by a high accumulation of snow during winter, the following spring might present a particularly thick bright layer. Moreover, it is reasonable to assume that seasons don't start and end exactly at the same time for different years. These considerations about the sedimentation rates and the duration of recurrent climatological phenomena (such as seasons) justify the assumption that there is misalignment in each of the three datasets $\{f_i^1\}_{i=1}^n$, $\{f_i^2\}_{i=1}^n$ and $\{f_i^3\}_{i=1}^n$. Therefore, in order to cluster weather, it is necessary to use k-mean alignment (see Section 1.4). Nevertheless, as previously explained, there are multiple ways to extend the k-means algorithm to the case of misalignend functional data. For example, in Arnqvist et al. (2016) and Pozzari (2017), the framework introduced in Section 1.4.1 was employed. However, due to the previously mentioned drawbacks, we decided to align and cluster functions using Algorithm 1.1 by switching to the SRVF representatives of the datasets.
As Figures 3.1 shows, lake Korttajärvi seems to provide more noisy signals. This is clearly shown in Figure 3.2, which highlights the wider range and higher dispersion of the average absolute value of the first derivative regarding lake Korttajärvi when compared with Kassjön and Nautajärvi. This could either be a specific feature of lake Korttajärvi or it could be due to a possible under-smoothing in the pre-processing step. Regardless of the reason, this characteristic could negatively influence the clustering procedure, since the SRVF representative of a function is directly linked to its first derivative. For this reason, the data from lake Korttajärvi have undergone a further smoothing step through cubic penalized B-spline basis functions[2].

[1]From now on, lake Kassjön will be referred to using index 1, lake Korttajärvi using index 2 and Nautajärvi using 3.
[2]The penalization parameter was set to 0.001

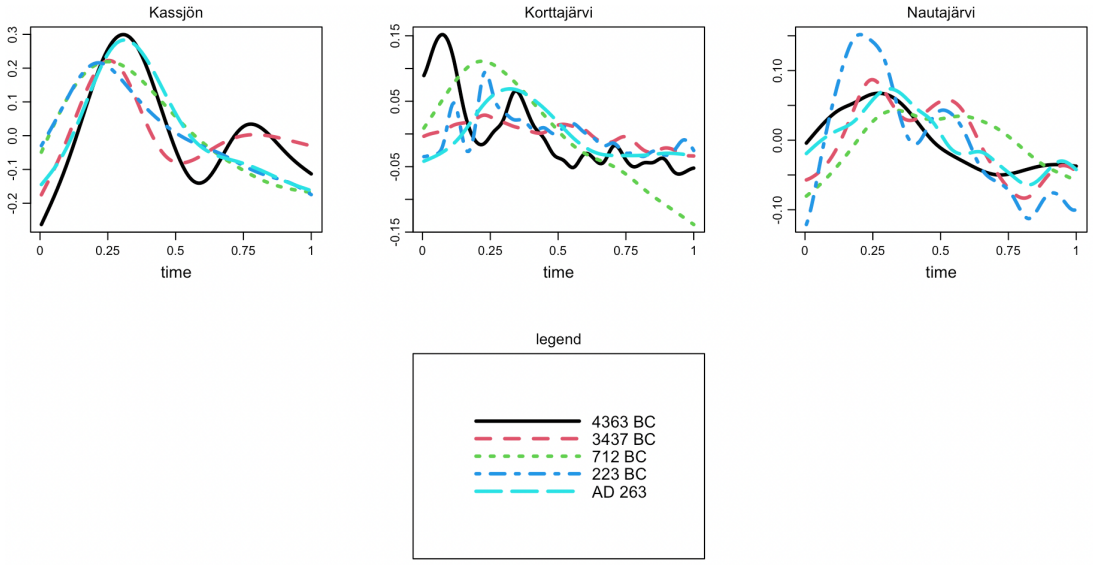


Figure 3.1: Sample of five years across the three lakes to show the variability of lake Korttajärvi's signals, when compared with the functions pertaining to the same years in the other two lakes. The color and line types match across the plots for the lakes.

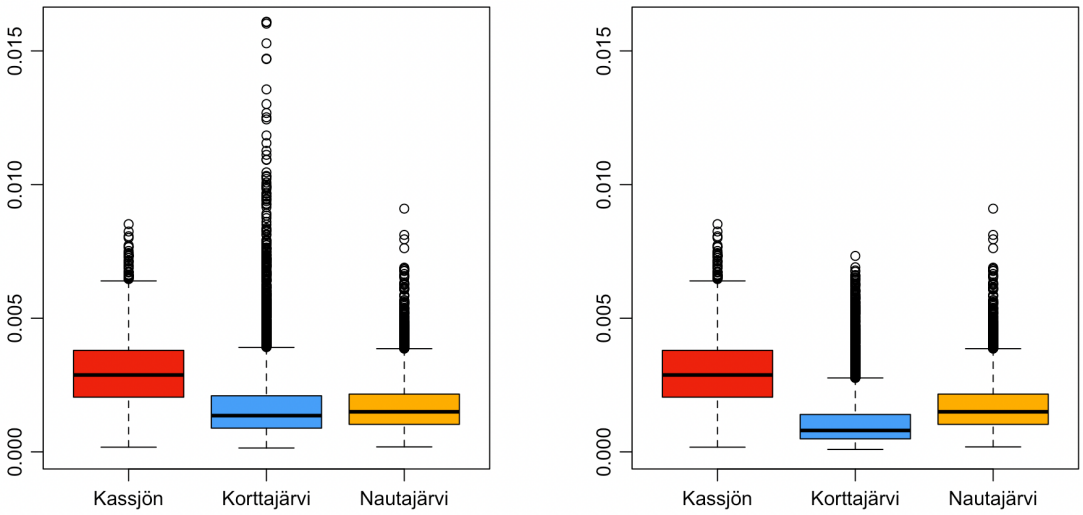


Figure 3.2: Boxplot of the cumulative absolute value of the first derivative $\left(\int_0^1 |\dot{f}(t)|dt\right)$ before (left) and after (right) the smoothing procedure applied to the signals from lake Kassjön.

3.1.1. Initial weather cluster templates

Algorithm 1.1 requires to initialize the weather cluster templates. Using a multi-start approach is computationally quite costly, thus suitable initial weather templates need to be properly chosen. For each signal $i \in \{1, 2, ..., n\}$ in each of the three datasets $\ell \in \{1\ (\text{Kassjön}), 2\ (\text{Korttajärvi}), 3\ (\text{Nautajärvi})\}$, the following five features were computed:

1. value of the highest peak: $\max_{t\in[0,1]} f_i^\ell$
2. L_1 displacement from the mean: $\int_0^1 |f_i^\ell(t) - \mu_\ell(t)|dt$ where $\mu_\ell(t) = \frac{1}{n}\sum_{i=1}^n f_i^\ell(t)$
3. integral of the absolute value of first derivative: $\int_0^1 |\dot{f}_i^\ell(t)|dt$
4. position of the highest peak: $\operatorname{argmax}_{t\in[0,1]} f_i^\ell$
5. Modified Band Depth (MBD) of function $f_i^\ell(t)$

Features 1, 2, 3 have been selected based on the fact that the intensity of the spring peak (which often coincides with the maximum value of the signal) and the velocity with which a function reaches the highest peak are believed to be important pieces of information carried by the data, since they are usually related to the most intense minerogenic inflow in the lake (see Ojala and Alenius (2005) and Ojala et al. (2008)). Moreover, Figure 3.3 shows that the majority of the signals reaches a peak during spring (i.e. roughly within the first quarter of the $[0, 1]$ interval), which in general is the season corresponding to the highest minerogenic accumulation due to the melting of the snow from the previous winter. However, a considerable number of functions exhibit their peak value later in the year. Based on this consideration, the location of the highest peak has been included in the features.

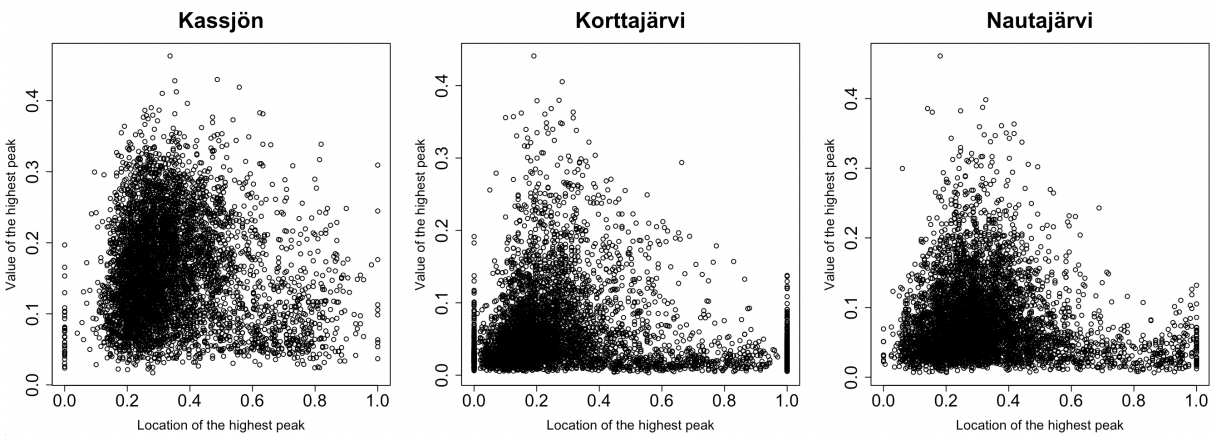


Figure 3.3: Scatterplot of the peak location vs peak intensity (normalized in $[0, 1]$). The signals having a peak exactly in 0 and 1 could be wrongly delimited years and are thus removed before clustering.

Finally, the Modified Band Depth has been included in the features to take into consideration the outlyingness of each function. This quantity measures how "central" a function is with respect to the rest, in terms of the fraction of time it is contained within a band constructed with a subset of the rest of the signals (see Ieva and Paganoni (2013) and

Lopez-Pintado et al. (2009)).
These five features work as a proxy for the information contained in the signals and allow to reduce the dimensionality of the functional data. The three datasets composed of the five features were then clustered through the "vanilla" k-means algorithm, each independently of the other two: indeed, considering its local nature, it was deemed inappropriate to cluster weather merging the information coming from all three lakes at the same time. To summarize, in order to obtain the k initial cluster templates for Algorithm 1.1, the following routine is employed for each lake $\ell \in \{1, 2, 3\}$:

1. Compute the five previously described features for each of the $n = 5271$ signals, thus moving from a functional to a multivariate framework
2. Cluster the multivariate dataset using the k-means algorithm
3. Compute the Tukey depth of each (multivariate) element in each cluster $j = 1, 2, 3, ..., k$
4. Set the initial template of cluster j for Algorithm 1.1 as the function corresponding to the "deepest year" in cluster j, i.e. the year associated with the deepest five-dimensional vector in cluster j (according to the Tukey depth)

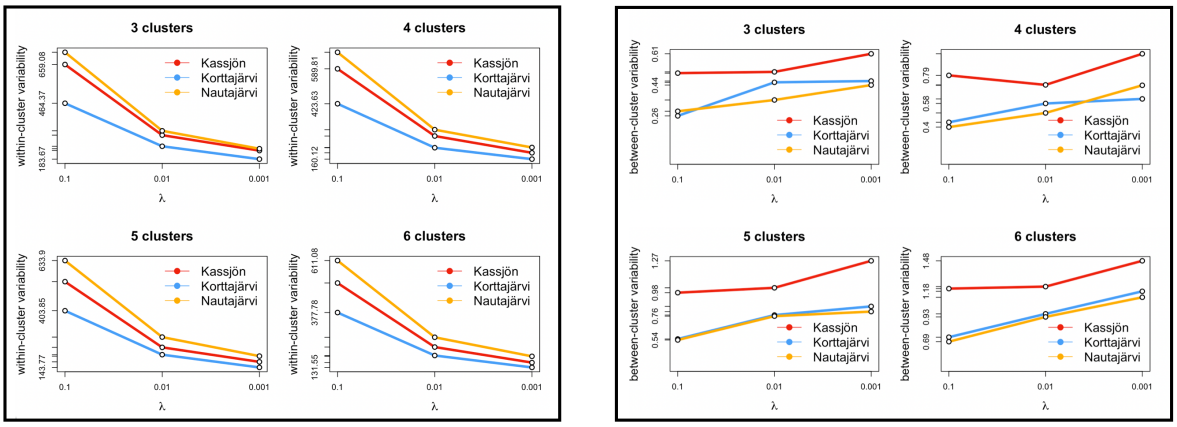
3.1.2. Mitigating the over-alignment

The authors of Srivastava et al. (2011) acknowledge that the tendency towards over-alignment in step 4 of Algorithm 1.1 represents the main limitation of the mathematical framework proposed in Section 1.4.2 (see Guo et al. (2022)). For this reason, the authors of the R library `fdasrvf` included a λ parameter in the registration step to penalize the over-alignment: the greater λ, the less alignment is allowed. Devising a metric to detect over-alignment is quite complicated, as the registration task depends greatly on the number of clusters and the number of functions to align. We clustered the years into three, four, five and six weather types and for each number of clusters we visually inspected the result of the alignment step, comparing the cases $\lambda = 0.1$, $\lambda = 0.01$ and $\lambda = 0.001$. Moreover, Figure 3.4 shows the between and within cluster variability as functions of the penalization parameter λ. The within- and between-cluster variability (W and B respectively) depend on the lake, the number of weather types k and parameter λ and are defined as

$$W = \sum_{j=1}^{k} \sum_{q \in \mathrm{W}j} \int_0^1 (q(t) - \psi_j(t))^2 dt$$

$$B = \sum_{j=1}^{k} \int_0^1 \left(\psi_j(t) - \frac{1}{n} \sum_{i=1}^{n} q_i(t) \right)^2 dt$$

where Wj indicates the set of registered functions (i.e. years) assigned to weather cluster j, ψ_j is the SRVF representative of cluster Wj and q_i is the SRVF representative of the i-th year aligned to its corresponding cluster, coeherently with the notation of Algorithm 1.1. Srivastava et al. (2011) propose to use W to shed light on the efficacy of the alignment step: intuitively, as λ decreases (i.e. when more alignment is allowed), the within cluster variability will decrease as well, since the signals will be registered more precisely to the corresponding cluster template. We decided to use $\lambda = 0.01$ for all lakes and for all clusters: indeed, decreasing λ from 0.1 to 0.01 leads approximately to a fifty percent decrease in within-cluster variability for almost all values of k (see Figure 3.4). On the other hand, the 10-fold decrease from 0.01 to 0.001 doesn't lead to significant changes in the weather labels. Figure 3.4b shows that, when compared to lake Kassjön, the Finnish lakes share similar between-variability scores for all values of λ and when k is at least four.



(a) (b)

Figure 3.4: Lake-wise within- (3.4a) and between- (3.4b) cluster variability as functions of $k \in \{3, 4, 5, 6\}$ and λ

3.1.3. Choosing the number of weather types k

Two techniques have been used in order to choose the number k of weather clusters in which to divide the years for each lake: the first consists in an elbow approach[3], while the second consists in trying to increase k as much as possible as long as the the new cluster contains information which is not already shared by the others[4].

Equation (3.1) indicates the average similarity of each signal to its corresponding weather

[3]This method is suggested in Sangalli et al. (2010)

[4]This approach is adopted in Pozzari (2017) and others

cluster as a function of k:

$$\rho_\ell(k) = 1 - \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (q_i^\ell - q_{\alpha_i}^\ell)^2 dt \qquad \forall \ell \in \{1,2,3\} \tag{3.1}$$

where q_i^ℓ is the SRVF representative of function i in lake ℓ and, according to the notation in Algorithm 1.1, α_i^ℓ represents the weather label assigned to the i-th year in the data set of lake ℓ. To be coherent with the metric used in the clustering algorithm, $\rho_\ell(k)$ is defined using the L_2 distance between q_i^ℓ and $q_{\alpha_i}^\ell$. As the range of the y-axis in Figure 3.5 clearly points out, the average similarity is high for all values of k[5], hence in order to choose the number of weather types it is necessary to investigate the presence of latent clusters for different values of k.

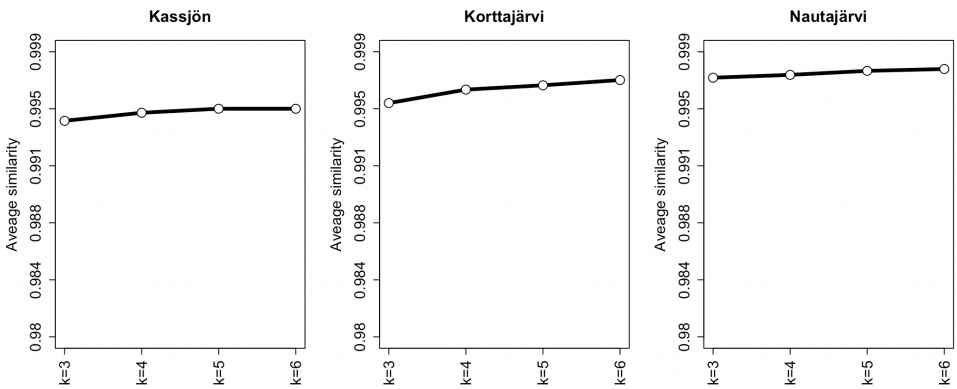


Figure 3.5: Average similarity for each lake $\rho_\ell = \rho_\ell(k)$ as a function of the number of weather clusters $k \in \{3,4,5,6\}$

Results: weather in lake Kassjön

Figure 3.6 is used to compare the results of the first-level clustering model when $k \in \{3$ (top), $4, 5, 6$ (bottom)$\}$. In particular, each row shows the content and cardinality (as a percentage in the title) of each weather cluster found when clustering the functions with a specific value of k. The gray lines are the functions contained in each weather cluster and aligned to the corresponding weather cluster representative (shown with a solid black line), which is defined as the point-wise average of the aligned functions in each cluster. When switching from $k = 3$ to $k = 4$, the new blue cluster is formed; then increasing k to five, the cyan cluster is created. When moving to $k = 6$, the newly formed purple weather type (W6) is similar to the blue one (W4) in terms of range and to the black one (W1) with respect to overall shape of the representative. The functions contained in any of these three weather types share the characteristic of having only one peak, which

[5]The same result holds if the alignment parameter λ is ten times higher

appears at different intensities. However, the same can be said when $k = 5$ considering the content of W1 and W4, hence it is possible to reach a similar partition of the years without adding the sixth cluster. For these reasons, we select $k = 5$ for lake Kassjön, since when k is increased to six one of the weather types is a latent cluster. Figure 3.7 emphasizes the similar content among weather clusters W4, W6 and W1 when $k = 6$.

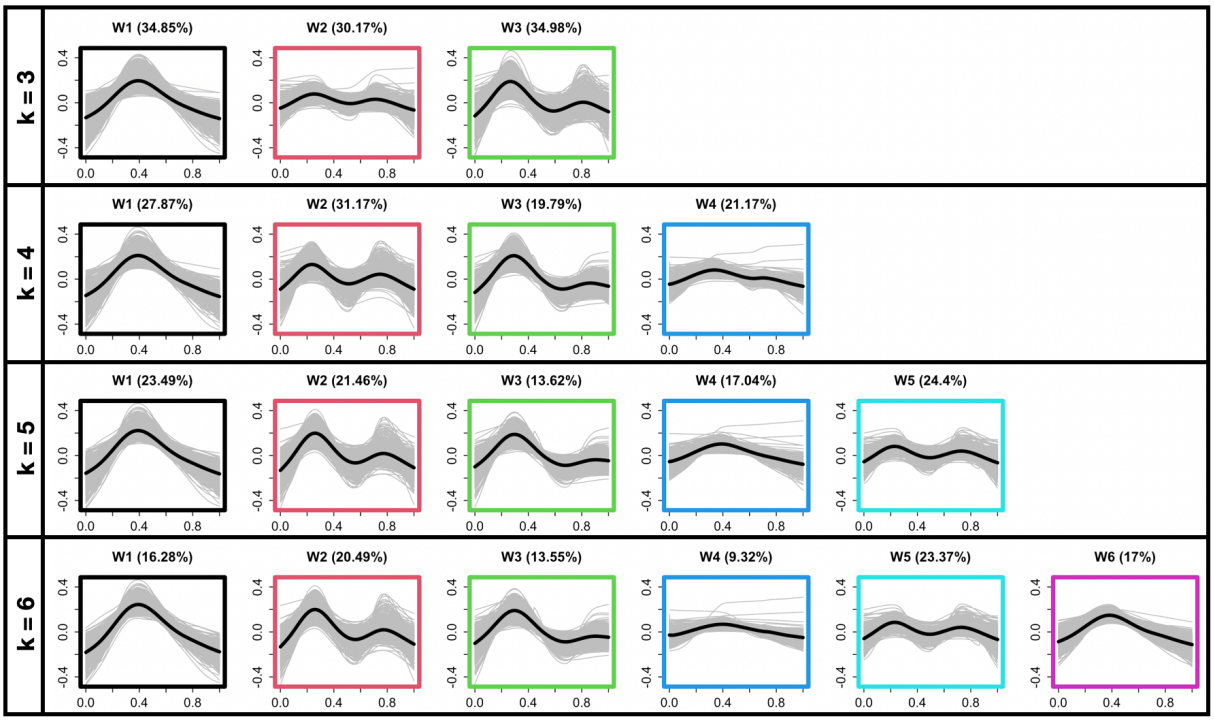


Figure 3.6: Weather types for lake Kassjön varying the number of clusters in $\{3, 4, 5, 6\}$. The percentage in the titles indicates the proportion of the 5271 years contained in each weather type. The x-axis is the interval $[0, 1]$.

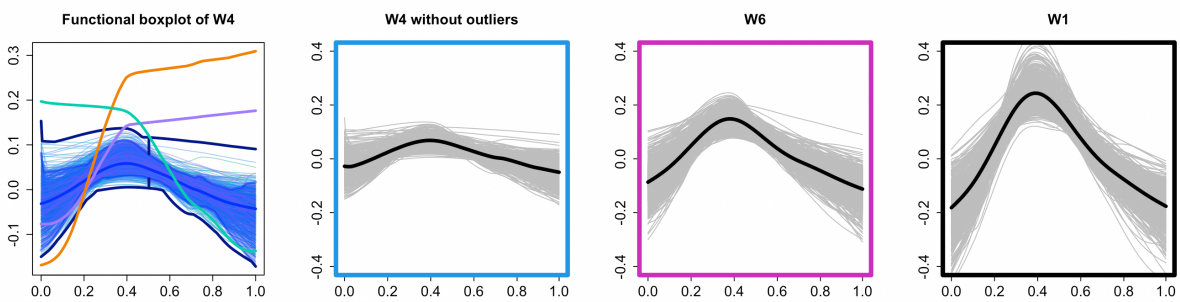


Figure 3.7: From left to right: functional boxplot of weather cluster W4, highlighting three outliers; weather cluster W4 without outliers; weather clusters W6 and W1, to compare with W4 (without outliers).

Results: weather in lake Korttajärvi

As explained in Section 3.1.3, each row of Figure 3.8 shows the output of the Algorithm 1.1 applied to lake Korttajärvi when $k \in \{3 \text{ (top)}, 4, 5, 6 \text{ (bottom)}\}$. Note that the y-axis are slightly different in each plot: indeed, the overall range of the signals obtained from the Finnish lakes is more narrow than the one from the Kassjön dataset. When k is increased from five to six the purple weather cluster W6 shares many similarities to W2. Moreover, the sum of the cardinalities of clusters W6 and W2 when $k = 6$ is comparable to the cardinality of W2 when $k = 5$, suggesting that when increasing the number of clusters from five to six, one weather type is roughly split into two. For these reasons, the number of weather types for Korttajärvi is set to five.

Results: weather in lake Nautajärvi

Figure 3.9 showcases the weather types of Nautajärvi varying $k \in \{3, 4, 5, 6\}$. Focusing on W4, as k increases the average intensity of the spring peak decreases and the behaviour around autumn and winter of the years assigned to this cluster becomes flatter and flatter. When $k = 6$, W4 and W3 share many similarities, as well as W4 and W6. In particular, Figure 3.10 shows that the SRVF representatives of W2 and W4 resulting from the clustering when $k = 6$ are comparable, while when k is set to five they differ in the last quarter of the x-axis. For these reasons, the number of weather types for Nautajärvi is set to five, since the addition of the sixth cluster does not seem to add new information.

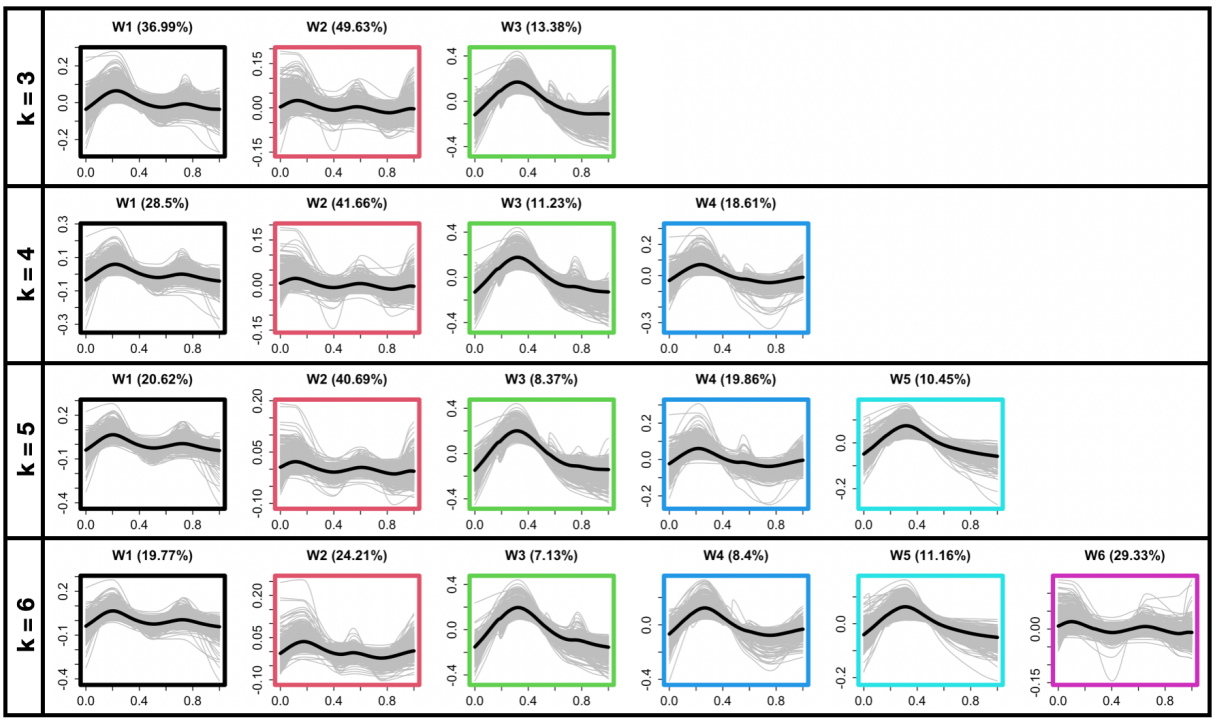


Figure 3.8: Weather types for lake Korttajärvi varying the number of clusters in $\{3, 4, 5, 6\}$. The percentage in the titles indicates the proportion of the 5271 years contained in each weather type. The x-axis is the interval $[0, 1]$.

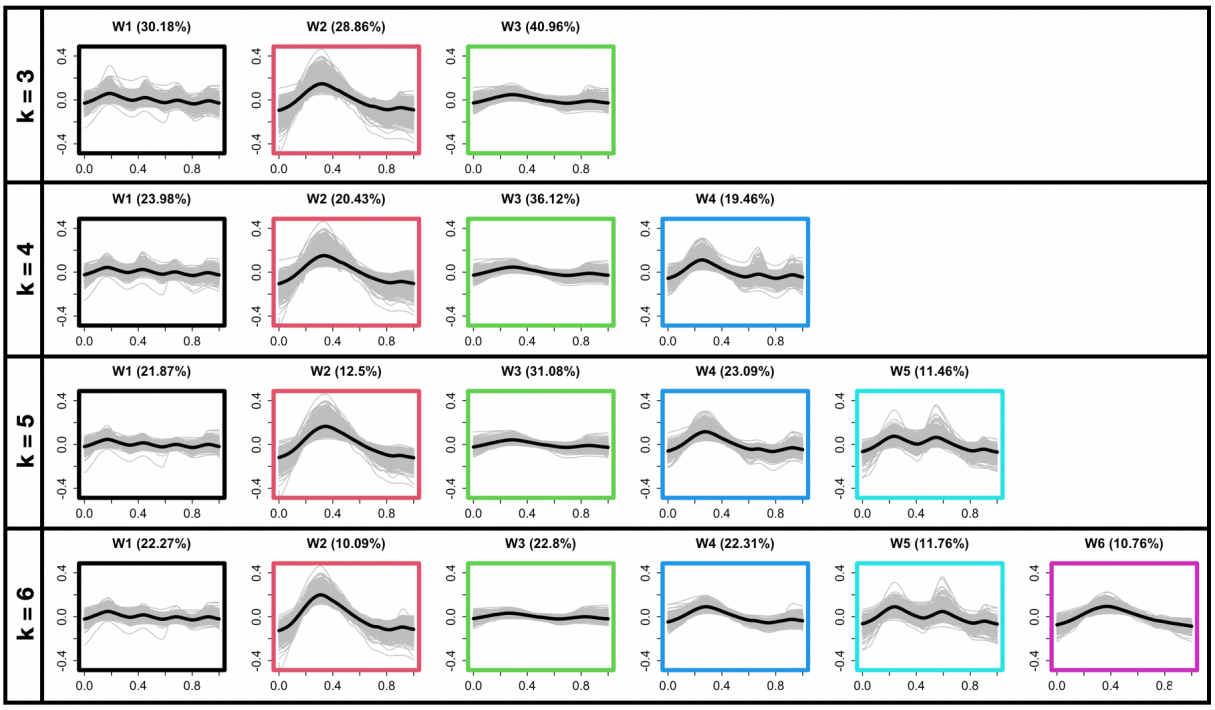


Figure 3.9: Weather types for lake Nautajärvi varying the number of clusters in $\{3, 4, 5, 6\}$. The percentage in the titles indicates the proportion of the 5271 years contained in each weather type.

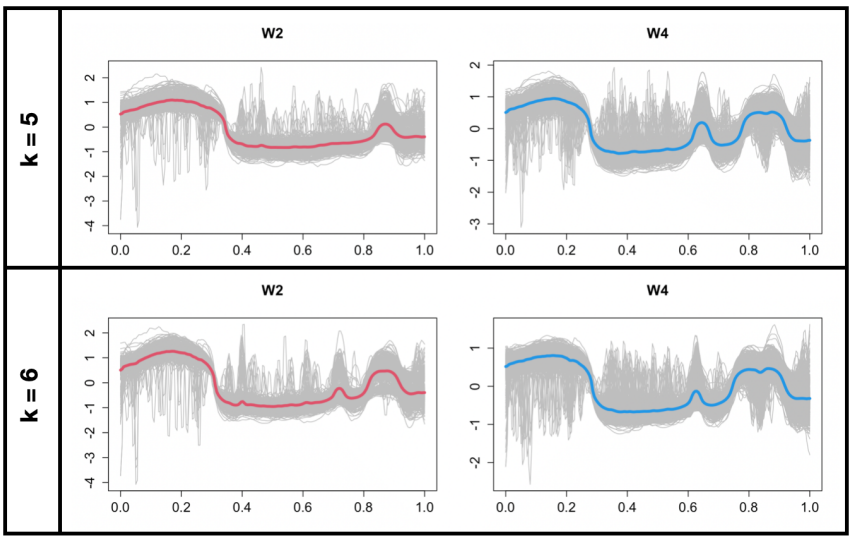


Figure 3.10: Plot of the SRVF representatives of W2 and W4 for lake Nautajärvi when $k = 5$ (top) and $k = 6$ (bottom)

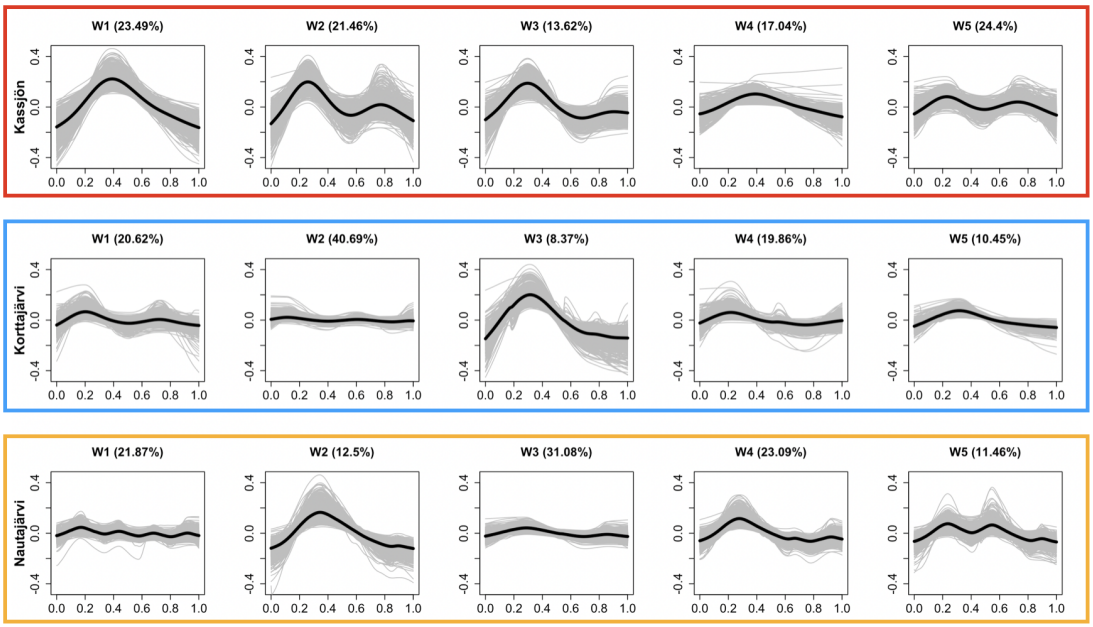


Figure 3.11: Content of the five weather types for each lake. The percentages in the titles indicate the fraction of the $n = 5271$ years contained in each cluster.

The final result of the first-level clustering is shown in Figure 3.11. Note that most weather templates don't match across different lakes: for example, weather type W1 for lake Kassjön does not correspond to W1 for Korttajärvi and Nautajärvi. After inspecting more thoroughly each cluster template, it is clear that some weather types are unique to some lakes. Consider, for example, the shape of W5 in lake Nautajärvi, uniquely characterized by two almost consecutive peaks: neither Korttajärvi nor Kassjön shows evidence

of a cluster template similar to W5 in Nautajärvi. Such differences are coherent with the geographically local nature of weather. Nevertheless, the lakes don't exhibit entirely different first-level clusters: consider, for example, the similarities between cluster W3 in Korttajärvi and W2 in Nautajärvi, or W3 in Kassjön and Korttajärvi. Generally, all lakes present three main categories of weather patterns: some are distinguished by having only one peak (such as W1 in Kassjön, W3 in Korttajärvi and W2 in Nautajärvi), others have a prominent spring peak followed by a more moderate second one during late summer or early autumn (such as W2 in Kassjön, W1 in Korttajärvi and W5 in Nautajärvi). Finally, there are some weathers which exhibit an almost flat behaviour (such as W4 in Kassjön, W2 in Korttajärvi and W3 in Nautajärvi). This entails that some of the five weather types found through Algorithm 1.1 can be viewed as "sub-weathers" of other clusters. With the help of experts, each of these tendencies can suit a specific interpretation of the years assigned to a specific weather type by each lake. For example, years belonging to W2 in Kassjön show a notable minerogenic inflow during spring, followed by a more moderate one during autumn. The latter might be a delayed effect of spring or it might have been caused by particularly heavy rains late in the year. The second peak could even indicate that an higher-than-average temperature during winter might have allowed sediments from the surrounding area to flow inside the lake.

As a result of the first-level clustering the 5271 years which compose the dataset are split into three sets of five weather clusters, one for each lake. Figure 3.12 shows the first level labels for each lake and clearly points out that such labels vary quite frequently from year to year. Indeed, it is not unusual for consecutive years to display different weather types, as established in the introduction to this work. A common assumption in the literature on this topic (see Arnqvist et al. (2016) and Abramowicz et al. (2019)) is that each sequence of yearly weather labels can be modeled using a Markov Chain with a suitable lake-specific transition matrix P_ℓ with $\ell \in \{1, 2, 3\}$. Each state of these Markov Chains represents the first-level label attributed to a specific year. Because of the Markov Property, working under this (potentially simplistic) assumption entails that the weather label assigned to year y only depends on the weather of year $y - 1$. From now on, the symbol $\{X_t^\ell\}$ will be used to represent the Markov Chain pertaining to lake $\ell \in \{1, 2, 3\}$, following the usual convention that index $\ell = 1$ refers to lake Kassjön, $\ell = 2$ to Korttajärvi and $\ell = 3$ to Nautajärvi. To be more explicit, $X_t^\ell \in \{1, 2, 3, 4, 5\}$ is the first level label assigned to year t according to lake ℓ. The 1-step transition probabilities estimated based on the weather labels shown in Figure 3.12 are reported in Table 3.1. Furthermore, Table 3.2 shows the estimated sojourn times in each of the five first-level labels and in each lake, proving the quickly varying nature of weather.

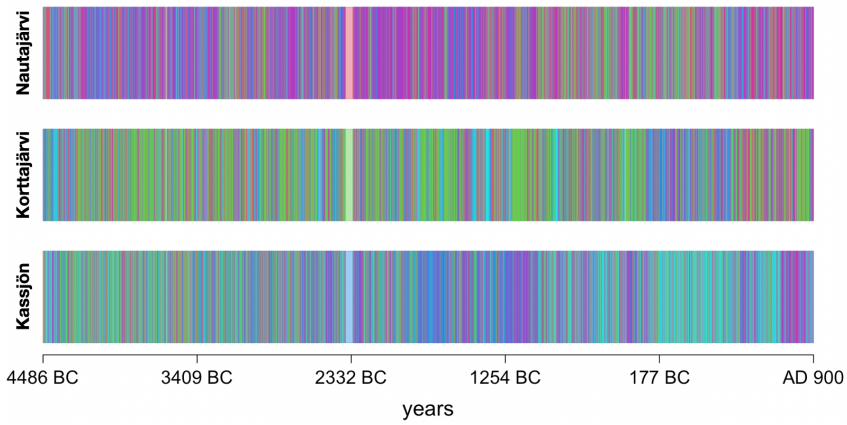


Figure 3.12: First-level labels for lake Kassjön (bottom), lake Korttajärvi (middle) and lake Nautajärvi (top). Each horizontal bar has been obtained by plotting a vertical line for each year, colored according to the weather label attributed to it by a specific lake.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	0.145	0.210	0.266	0.237	0.142
$i=2$	0.140	0.294	0.204	0.240	0.123
$i=3$	0.121	0.177	0.365	0.120	0.218
$i=4$	0.150	0.232	0.120	0.359	0.138
$i=5$	0.128	0.148	0.274	0.220	0.229

(a) Transition matrix for lake Kassjön (P_1)

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	0.223	0.443	0.172	0.054	0.109
$i=2$	0.174	0.616	0.131	0.021	0.058
$i=3$	0.222	0.202	0.331	0.096	0.149
$i=4$	0.152	0.084	0.238	0.410	0.116
$i=5$	0.240	0.189	0.290	0.102	0.180

(b) Transition matrix for lake Korttajärvi (P_2)

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	0.275	0.141	0.192	0.128	0.263
$i=2$	0.314	0.316	0.068	0.162	0.140
$i=3$	0.174	0.053	0.312	0.094	0.367
$i=4$	0.293	0.185	0.171	0.131	0.220
$i=5$	0.181	0.065	0.251	0.093	0.409

(c) Transition matrix for lake Nautajärvi (P_3)

Table 3.1: Estimated transition probabilities for the Markov Chains modeling the first level labels. In particular, $P_\ell(i,j) = \mathbb{P}(X^\ell_{t+1} = j | X^\ell_t = i)$ with $i,j = 1,2,3,4,5$ and $\ell = 1,2,3$.

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
Lake Kassjön	1.170	1.416	1.574	1.561	1.298
Lake Korttajärvi	1.286	2.603	1.495	1.697	1.219
Lake Nautajärvi	1.380	1.461	1.454	1.150	1.692

Table 3.2: Estimated sojourn times in each state (weather label) according to each lake

Arnqvist et al. (2016) propose to use a permutational test to investigate the time dynamics of the first level labels. For each lake $\ell \in \{1, 2, 3\}$ and for each couple of weather types i, $j \in \{1, 2, 3, 4, 5\}$, the normalized 1-step transition probability Q^{ℓ}_{ij} is computed:

$$Q^{\ell}_{ij} = \frac{\mathbb{P}(X^{\ell}_{t+1} = j | X^{\ell}_{t} = i)}{\mathbb{P}(X^{\ell}_{t+1} = j)} \tag{3.2}$$

Under the null hypothesis that the weather labels are randomly distributed across the years, Q^{ℓ}_{ij} would be equal to one since there would be no temporal structure between consecutive weather labels. Under this null hypothesis, the weather labels for each lake have been permuted one hundred thousand times and, for each permutation, the matrix Q^{ℓ}_{ij} and the pvalues for the corresponding test were computed for all three lakes and for all i, j. The result of these 25 tests for each lake are presented in Table 3.3: at a significance level of 0.01, there is evidence of the existence of a temporal structure in the weather labels for each lake and for each weather transition.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	1.06	0.98	1.09	1.01	0.83
$i=2$	1.03	1.37*	0.84*	1.02	0.72*
$i=3$	0.89	0.82*	1.50*	0.51*	1.28*
$i=4$	1.10	1.08	0.49*	1.53*	0.81*
$i=5$	0.94	0.69*	1.12	0.94	1.35*

(a) Q_{ij} matrix for Kassjön

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	1.12	1.09*	0.83*	0.64*	1.04
$i=2$	0.88*	1.51*	0.64*	0.25*	0.56*
$i=3$	1.12	0.49*	1.61*	1.14	1.43*
$i=4$	0.76	0.21*	1.15	4.91*	1.11
$i=5$	1.21	0.46*	1.41*	1.21	1.72*

(b) Q_{ij} matrix for Korttajärvi

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	1.19*	1.13	0.88	1.12	0.85*
$i=2$	1.36*	2.52*	0.31*	1.42*	0.45*
$i=3$	0.76*	0.42*	1.43*	0.82	1.18*
$i=4$	1.27*	1.48*	0.78*	1.14	0.71*
$i=5$	0.79*	0.52*	1.15*	0.82*	1.32*

(c) Q_{ij} matrix for Nautajärvi

Table 3.3: Normalized 1-step transition probabilities for lake Kassjön (3.3a), Korttajärvi (3.3b) and Nautajärvi (3.3c). The figures with a star are different from 1 at a significance level equal to 0.01, even when applying the Bonferroni correction.

3.2. Second-level clustering

The first-level clustering was used to reconstruct the weather in each lake taking into consideration the misalignment in each dataset. The role of the second-level clustering is to use this information to cluster climate, managing the temporal dependence of the functions coming from each lake: it is reasonable to assume that the climate of year y is the same or at least similar to the one of the adjacent years, hence the temporal dependence. It is important to stress that while the first-level clustering provided three sets of weather labels (one for each lake), the result of the second-level clustering is one single set of climate labels, shared by all lakes.

3.2.1. Bagging Voronoi algorithm

To take the temporal dependence into considerations, the second-level clustering is based on the Bagging Voronoi (BV) method detailed in Algorithm 3.1. As a result of the first-level clustering, each year has been assigned three labels, one for each lake. Therefore, the Bagging Voronoi algorithm is applied to the set of triplets of yearly weather labels. BV is a flexible method, used in previous papers to cluster spatially-dependent functional data defined on a multidimensional lattice (e.g. Secchi et al. (2013), Menafoglio and Secchi (2017)). In this section, we will explain the main steps of this model, specifically tailored for our goal.

In short, N years out of the 5271 in the dataset are randomly selected and called "nuclei". Then, the 5271 years are divided into non-overlapping cells, each containing on average $L = [5271/N]$ elements, according to which is the closest nucleus. The collection of these cells is called Voronoi tessellation. The next step consists in representing the years within each cell with a suitable representative, defined ad hoc for the task at hand. The set of N cell representatives is then partitioned into a suitable number of groups using a given clustering model. In the end, each year will be assigned the label of its corresponding cell. This entails that on average L consecutive years obtain the same climate label. Furthermore, the clustering is not performed on the data directly, but on an auxiliary dataset (see Menafoglio and Secchi (2017)) constructed with the cell representatives. This is the reason why the Bagging Voronoi algorithm is used to take the temporal dependence of adjacent years into consideration. This procedure is run B times, to cope with the random choice for the nuclei.

The extensive flexibility of the Bagging Voronoi algorithm stands out considering the fact that it imposes minimal constraints on how to represent each Voronoi cell and on the choice for the model used to cluster such cell representatives. What follows is the detailed

description of the Bagging Voronoi model applied to $\{\boldsymbol{g}_i\}_{i=1}^n$, where $\boldsymbol{g}_i = [g_i^1, g_i^2, g_i^3]$ is a three-dimensional vector: $g_i^\ell \in \{1, 2, 3, 4, 5\}$ is the first-level (weather) label assigned to the i-th year according to lake ℓ.

Algorithm 3.1 Bagging Voronoi tessellation

Let $\{\boldsymbol{g}_i\}_{i=1}^n$ be the three-dimensional weather labels on the lattice of $n = 5271$ years
Fix number of clusters K, a temporal resolution parameter L, number of bootstrap iterations B
Set $S_0 = \{-4486, -4485, ..., 900\}$ as the one-dimensional lattice of years
Define the number of nuclei $N = \left[\frac{n}{L}\right]$
Bootstrap step:
for $b = 1, 2, ..., B$ **do**
  Randomly generate a set of N nuclei $\Phi_N^b = \{Z_1^b, Z_2^b, ..., Z_N^b\}$ such that $Z_c^b \overset{i.i.d.}{\sim} \mathcal{U}(S_0) \quad \forall c = 1, 2, ..., N$
  Obtain a random Voronoi tessellation of S_0 in N cells, called $\{V_c^b := V(Z_c^b|\Phi_N^b)\}_{c=1}^N$, by assigning each year in S_0 to the nearest nucleus Z_c^b according to a suitable distance (e.g. L_1 or L_2)
  for $c = 1, 2, ..., N$ **do**
    Compute $\bar{\boldsymbol{g}}_c^b$, a *local* representative of the data in the c-th cell
  end for
  Using a suitable model $\mathcal{M}$, cluster the N cell representatives
  for $i = 1, 2, ..., 5271$ **do**
    Assign to the i-th year the label of its cell
  end for
  Return the labels of the b-th bootstrap iteration, which create a partition of the n years into K clusters $\{C_j^b\}_{j=1}^K$
end for
Aggregation step:
Using the framework proposed in Stephens (2000), relabel the cluster assignments $\{\{C_j^b\}_{j=1}^K\}_{b=1}^B$
for $i = 1, 2, ..., 5271$ **do**
  Compute $\pi_i^j = \#\{b \in \{1, 2, ..., B\} : i \in C_j^b\}/B \quad \forall j = 1, 2, ..., K$
  Assign to year i the label $j^* = \underset{j=1,2,...,K}{\operatorname{argmax}}\ \pi_i^j$, i.e. its most frequently assigned label throughout the B bootstrap iterations
  Compute the *entropy* of the i-th year as $\eta_i = -\left(\sum_{j=1}^K \pi_i^j \log \pi_i^j\right)$
end for

The goal of the bootstrap step is to construct an auxiliary dataset distinguished by a lower temporal dependence with respect to the original one, as each datum serves as a substitute for L consecutive years on average. As L increases, the cells become wider, hence the temporal dependence of their representatives will become more negligible. Moreover, as L increases, the number of consecutive years that on average acquire the same (climate) label increases as well, making the transition from one climate type to another quite abrupt. On the other hand, as L decreases, the climate clusters will change more gradually in time. These considerations point out that the temporal resolution parameter L requires careful tuning.
Since each bootstrap iteration is based on different cells, it is not guaranteed that the climate labels are preserved: this is why the labels assigned to the years across different bootstrap iterations undergo a relabeling procedure according to Stephens (2000). Finally, after the label switching step, each year is given its most-frequently assigned climate label and the yearly entropy is computed using the formula at step 21 of Algorithm 3.1: the higher η_i, the more "unsure" the model is about the climate type assigned to year i. The average entropy $\bar{\eta} = \bar{\eta}(K, L) = \sum_{i=1}^{n} \eta_i / n$ can be used as an indicator of overall robustness of the clustering model for a specific choice of temporal resolution L and number of climate types K. In order to compare different $\bar{\eta}(K, L)$ in terms of K, a normalized entropy can be defined as $\bar{\eta}(K, L)/K$.

3.2.2. Cell representative

Given that $\{\boldsymbol{g}_i^b = [g_i^{b,1}, g_i^{b,2}, g_i^{b,3}]\}_{i=1}^n$ is the set of triplets of yearly weather labels and following the definition of climate as the occurrence of a specific weather pattern over time, the representative of each cell is a three-dimensional object. The first component of a cell representative is the frequency of occurrence of each weather type within the cell according to the first-level labels of Kassjön, the second component according to the weather labels of Korttajärvi and the third of Nautajärvi. To simplify the notation in this subsection, the b index will be omitted from the symbol for the cell representative $\bar{\boldsymbol{g}}_{cell}^b$, keeping in mind that each bootstrap iteration provides a different set of cells (and thus cell representatives). Defining $I := \{1, 2, 3, 4, ..., n\}$ as the indices of the years in S_0 and k the number of first-level labels for each lake, the cell representative $\bar{\boldsymbol{g}}_{cell} = [\bar{g}_{cell}^1, \bar{g}_{cell}^2, \bar{g}_{cell}^3]$ is defined as

$$\bar{\boldsymbol{g}}_{cell} = \begin{bmatrix} \left\{ \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^1_i = 1\}|}{|V^b_{cell}|}, \quad \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^1_i = 2\}|}{|V^b_{cell}|}, \ldots, \quad \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^1_i = k\}|}{|V^b_{cell}|} \right\} \\ \left\{ \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^2_i = 1\}|}{|V^b_{cell}|}, \quad \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^2_i = 2\}|}{|V^b_{cell}|}, \ldots, \quad \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^2_i = k\}|}{|V^b_{cell}|} \right\} \\ \left\{ \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^3_i = 1\}|}{|V^b_{cell}|}, \quad \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^3_i = 2\}|}{|V^b_{cell}|}, \ldots, \quad \frac{|\{i \in I \,:\, S_0[i] \in V^b_{cell} \text{ and } g^3_i = k\}|}{|V^b_{cell}|} \right\} \end{bmatrix}$$

Following the usual convention throughout this thesis, indices 1, 2 and 3 refer to lake Kassjön, Korttajärvi and Nautajärvi, respectively. In general, the number of first-level (weather) labels k might be different for each lake, however, since we found that $k = 5$ for all first-level labels (see Section 3.1.3), we will present the model working with the same k for all lakes.

3.2.3. Clustering model $\mathcal{M}$

The clustering model $\mathcal{M}$ is a crucial step in the Bagging Voronoi algorithm: $\mathcal{M}$ is responsible for assigning the climate labels to the N cells and thus to the n years for each iteration $b = 1, 2, ..., B$. In particular, this clustering model should merge the information contained in the three lakes. In order to achieve this goal, we devised model $\mathcal{M}$ as an iterative algorithm based on the multivariate k-means method applied to the auxiliary dataset made of the N cell representatives $\{\bar{\boldsymbol{g}}^b_c\}_{c=1}^N$ at the b-th bootstrap iteration. From now on, we will omit the b index and assume implicitly that the clustering needs to be performed B times. The overall outline of our proposed cell-clustering model $\mathcal{M}$ applied to the data set of N cell representatives is as follows:

Algorithm 3.2 Second-level clustering

```
Fix number of clusters K, maximum number of iterations MaxItr, threshold ϵ
Initialize the K cluster templates
for it = 1, 2, ..., MaxItr do
    Assign each cell to the closest cluster
    Update the cluster templates
    if distance between new and old cluster representatives is smaller than ϵ then
        Stop
    end if
    it = it + 1
end for
```

Essentially, model 3.2 is an iterative procedure which differs from the standard k-means algorithm regarding the choice for the clusters' templates and the metric used to measure

the distance between each cell and each cluster. Indeed, while within the conventional framework of the "vanilla" k-means algorithm the dataset usually consists of multivariate vectors, in the context of this thesis the statistical units to partition in K clusters are represented by *triplets of distributions*.

Choice for cluster template

When applying the k-means algorithm to a p-dimensional dataset, each cluster template is computed as the average of the p-dimensional data assigned to the cluster itself at a specific iteration. On the contrary, in the version of the k-means developed in this thesis, each climate cluster representative is defined in the same way as the cell representatives as explained in Section 3.2.2, i.e. as the *triplet* of lake-wise distributions of weather labels inside each *cluster*. The reason behind the use of three separate distributions to describe each cluster is to allow each lake to be properly taken into consideration: indeed, if the cluster representative didn't separate the information coming from the three lakes, it would be complicated to reconstruct the agreement of each lake with the final climate clusters.
Finally, the templates for the climate types are updated using the conventional updating step of k-means applied one-lake-at-a-time. This means that each of the three distributions describing a climate cluster is updated as the average of the corresponding distributions representing the cells within the cluster.

Soft-majority vote

The main difference between conventional k-means and the algorithm we propose is the notion of distance between a cell and a cluster representative. Since the cells and clusters are represented through three-dimensional distributions, in order to quantify how close each cell is to the K clusters we propose to use a weighted distance between each of the three distributions of the cell and the cluster. To explain the role of the weighted distance, consider the toy example in Figure 3.13, which shows a specific iteration of the algorithm assuming $K = 4$. The four cluster templates CL1, CL2, CL3 and CL4 are the result of the template updating steps performed throughout the previous iterations of the model. For example, the first (red-colored) "component" of the cluster template for CL1 represents how frequently each weather type of lake Kassjön occurs within climate cluster CL1 according to the current iteration. In this example, out of all the years assigned to CL1 at this iteration, approximately 40% have been assigned to weather type W4 as a result of the first level clustering procedure, roughly 30% to W3, 20% to W2 and the remaining years belong to weather types W1 and W5.

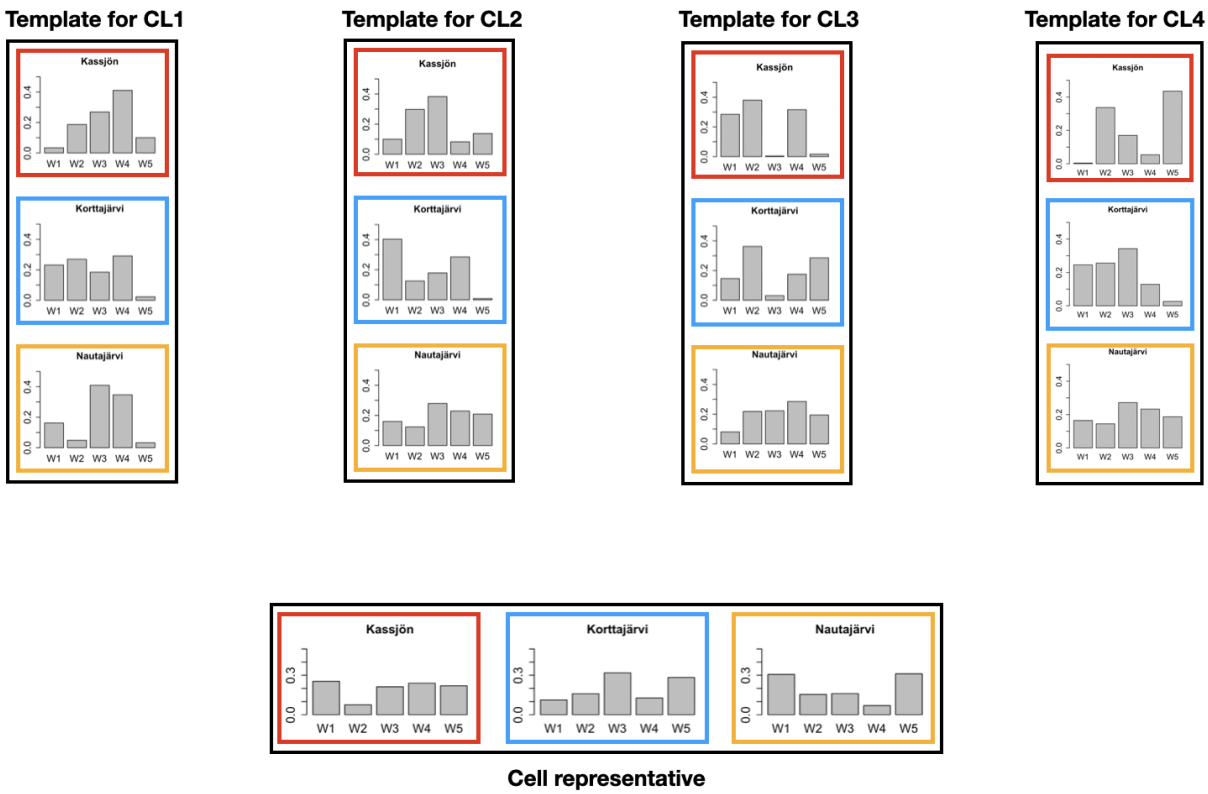


Figure 3.13: Toy example of second-level clustering at a specific iteration of model $\mathcal{M}$ with number of climate types $K = 4$, for simplicity. The histograms coloured in red refer to the cell or cluster representatives for Kassjön, the blue ones to Korttajärvi and the orange ones to Nautajärvi. For example, the red histogram of the template for CL1 indicates the normalized number of times in which each weather type of Kassjön appears inside the years which have been assigned to CL1 at the current iteration.

Just as in the standard k-means, the cell will be assigned a label corresponding to the closest cluster, hence it is necessary to define a distance between a cell representative $\bar{\boldsymbol{g}}_c = [\bar{g}_c^1, \bar{g}_c^2, \bar{g}_c^3]$ $\forall c = 1, 2, ..., N$ and a cluster representative $\hat{\boldsymbol{g}}_j = [\hat{g}_j^1, \hat{g}_j^2, \hat{g}_j^3]$ $\forall j = 1, 2, ..., K$. Such distance δ is defined as

$$\delta(\bar{\boldsymbol{g}}_c, \hat{\boldsymbol{g}}_j) = w_c^1 d(\bar{g}_c^1, \hat{g}_j^1) + w_c^2 d(\bar{g}_c^2, \hat{g}_j^2) + w_c^3 d(\bar{g}_c^3, \hat{g}_j^3) \tag{3.3}$$

where the first term refers to the distance d between the cell and cluster representative of Kassjön, the second addend refers to Korttajärvi and the third one to Nautajärvi. Furthermore, w_c^1, w_c^2 and w_c^3 are suitable weights pertaining to lake Kassjön, Korttajärvi and Nautajärvi respectively (more on the choice for the weights later). Finally, metric d measures the distance between two probability density functions. As pointed out in Abramowicz et al. (2019), d could be chosen as a weighted version of the L_2 distance or as a proper distance between distributions, such as total variation distance, Wasser-

stein distance or others. After some experiments we chose to use the Jensen-Shannon divergence,

$$JS(p_1, p_2) = H\left(\frac{p_1 + p_2}{2}\right) - \frac{1}{2}H(p_1) - \frac{1}{2}H(p_2) \tag{3.4}$$

where $p_1,\ p_2$ are two distributions and H is the Shannon entropy function. The Jensen-Shannon divergence is based on the Kullback–Leibler divergence, it is symmetric and doesn't require the assumption of absolute continuity of p_1 with respect to p_2 (see Lin (1991) for more details).
The cell could be clustered to j^*, where

$$j^* = \underset{j=1,2,\ldots,K}{\operatorname{argmin}}\ \delta(\bar{\boldsymbol{g}}_c, \hat{\boldsymbol{g}}_j) \tag{3.5}$$

However, it is possible to further exploit the information coming from each lake in the following way. Before assigning a label to a cell according to all lakes, we first cluster the cell to a climate type according to the opinion of each lake, independently of the other two. Thus, we obtain three *candidate labels* for the cell, called j_1^* proposed according to lake Kassjön's dataset, j_2^* proposed according to Korttajärvi's one, and j_3^* according to Nautajärvi's signals. More precisely,

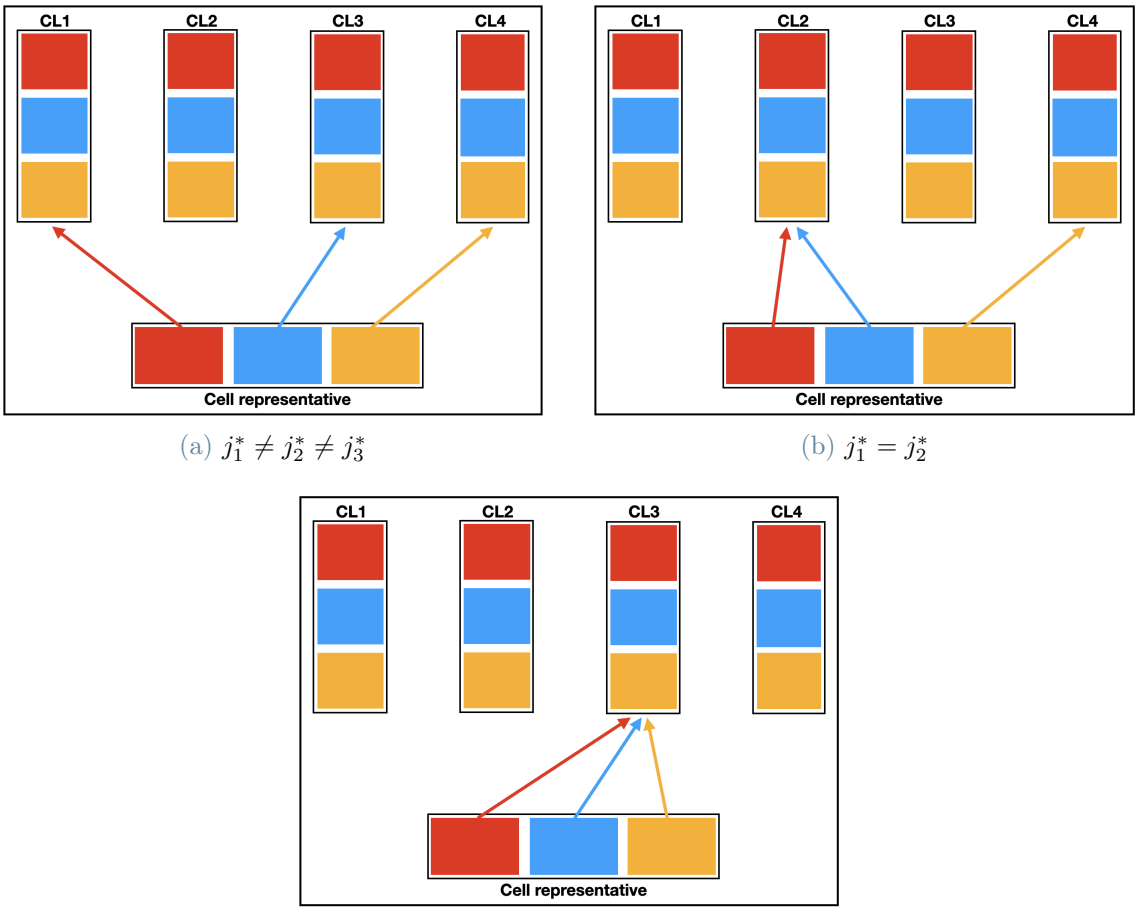
$$\begin{aligned}
j_1^* &= \underset{j=1,2,\ldots,K}{\operatorname{argmin}}\ d(\bar{g}_c^1, \hat{g}_j^1) && \text{[Candidate cell label according to Kassjön]} \\
j_2^* &= \underset{j=1,2,\ldots,K}{\operatorname{argmin}}\ d(\bar{g}_c^2, \hat{g}_j^2) && \text{[Candidate cell label according to Korttajärvi]} \\
j_3^* &= \underset{j=1,2,\ldots,K}{\operatorname{argmin}}\ d(\bar{g}_c^3, \hat{g}_j^3) && \text{[Candidate cell label according to Nautajärvi]}
\end{aligned} \tag{3.6}$$

The final label j^* assigned to the cell is

$$j^* = \underset{j \in \{j_1^*,\ j_2^*,\ j_3^*\}}{\operatorname{argmin}}\ \delta(\bar{\boldsymbol{g}}_c, \hat{\boldsymbol{g}}_j) \tag{3.7}$$

Differently from (3.5), according to (3.7) it is guaranteed that at least one lake fully agrees with the cell assignment, since only the three candidate labels are explored. We named the routine of clustering climate using model $\mathcal{M}$ based on (3.7) *soft-majority vote*. Figure 3.14 shows three scenarios: in 3.14a all lakes disagree, since each of them would label the cell differently from the other two. Hence, whichever the climate type attributed to the current cell, the assignment will coincide with the candidate label proposed only by one lake out of three. For future reference, the cell assignments of this type are called "assignments of the first type". In Figure 3.14b, two out of three lakes agree. If the

winning candidate label is the one suggested by the two lakes, this assignment will be called "assignment of the second type". Finally, the best case scenario is shown in Figure 3.14c: all lakes agree, hence the name "assignment of the third type". In this latter case, the cell is guaranteed to be clustered according to the common candidate label. These partition of the cell assignments into three types will later be used to shed light on the agreement of each lake with respect to the final climate labels.



(a) $j_1^* \neq j_2^* \neq j_3^*$ (b) $j_1^* = j_2^*$

(c) $j_1^* = j_2^* = j_3^*$

Figure 3.14: Examples of different types of cell assignment: first-type (3.14a), second-type (3.14b) and third-type (3.14c). The coloured arrows point to the candidate labels proposed by each lake [red: Kassjön; blue: Korttajärvi; orange: Nautajärvi]. For example, in 3.14a, Kassjön would assign the cell to cluster CL1, Korttajärvi to CL3 and Nautajärvi to CL4.

Choice for weights

The weights in equation (3.7) are used to quantify the role of each lake in the cell-labeling process. Therefore, in the most general case, such weights are not only lake-, but also cell-specific. In addition to their role in accentuating or mitigating the impact of each lake, these weights could potentially serve as a means of incorporating the perspectives of

domain experts into the model. For example, if for any reason it were deemed appropriate to distrust the data coming from one lake, then it would be reasonable to set smaller weights for that lake. In Chapter 2 it was noted that the signals from lake Korttajärvi are particularly noisy. Such noise could be due to inflows coming from two basins located in the geographical area nearby lake Korttajärvi. The experts' opinion could help in quantifying the degree of distrust to be attributed to lake Korttajärvi in the cell-labeling process.

We propose an alternative way of choosing the lake- and cell-specific weights in (3.7), using an a priori estimate about the degree of uncertainty of the model with respect to each year's climate label. The steps taken to achieve this are the following:

1. For each lake ℓ, reduce the dimensionality of the data set passing from a functional setting to a multivariate one, by using as features:
   - $\max_{t\in[0,1]} f_i^\ell \quad \forall i = 1, 2, ..., n$
   - $\int_0^1 |f_i^\ell(t) - \mu_\ell(t)|dt$ where $\mu_\ell(t) = \frac{1}{n}\sum_{i=1}^n f_i^\ell(t) \quad \forall i = 1, 2, ..., n$
   - $\int_0^1 |\dot{f}_i^\ell(t)|dt \quad \forall i = 1, 2, ..., n$
2. Compute the first two principal components of each of the three data sets made with such features.
3. Apply Algorithm 3.1 to the bivariate data sets constructed with the first two scores resulting from the PCA on the described features. In particular, represent each cell with a two-dimensional vector, which corresponds to a weighted average of the two-dimensional arrays containing the first and second principal components of the years within the cell. The weight assigned to each element correspond to its Mahalanobis depth with respect to the other vectors in the cell: in this way each local cell representative is mostly influenced by the deepest points. As clustering model $\mathcal{M}$ we used the vanilla k-means.
4. For each lake ℓ and for each year i, compute the entropy η_i according to step 21 of Algorithm 3.1. In particular, we obtain:

$$\eta_i^\ell = -\left(\sum_{j=1}^K \pi_i^j \log \pi_i^j\right)$$

   for each lake $\ell \in \{1, 2, 3\}$ and year $i \in \{1, 2, ..., n\}$. As explained before, π_i^j represents the frequency of times in which year i has been assigned to climate cluster j out of the B bootstrap iterations.

Note that this routine treats each lake independently of the other two, thus it allows to obtain three entropy function, each quantifying a priori (i.e. before computing the proper climate labels) the degree of uncertainty of each lake in each year and with respect to a specific choice for the number of climate clusters K and temporal resolution L. For example, if one were to investigate the presence of four distinct climate types assuming that on average each cell contains three hundred years, then it would be necessary to first compute the three a priori entropy functions for $K = 4$ and $L = 300$.
The weights in equation (3.7) are computed as the inverse of the average a priori entropy of the years contained in the cell. In this way, the lake which is estimated to be the most unsure regarding the years within a cell (i.e. the lake which has the highest average entropy), will be assigned a lower weight. For example, consider Figure 3.15: on the bottom the differently colored intervals represent the eighteen cells resulting from the choice of $L = 300$. On the top, the three colored lines represent the yearly entropy functions computed through the procedure explained in this section, when $K = 4$ and the number of bootstrap iteration B is equal to 500. The average yearly entropies within the gray-colored cell spanning from 2000 BC to 1485 BC are 0.495 for Kassjön, 0.544 for Nautajärvi and 1.077 for Korttajärvi. The weights are $2.022 = 1/0.495$ for Kassjön, $1.839 = 1/0.544$ for Nautajärvi and $0.928 = 1/1.077$ for Korttajärvi. In this cell, this a priori analysis shows that Korttajärvi has the highest entropy, hence it is reasonable to underweight the candidate label of the gray cell proposed by this lake.

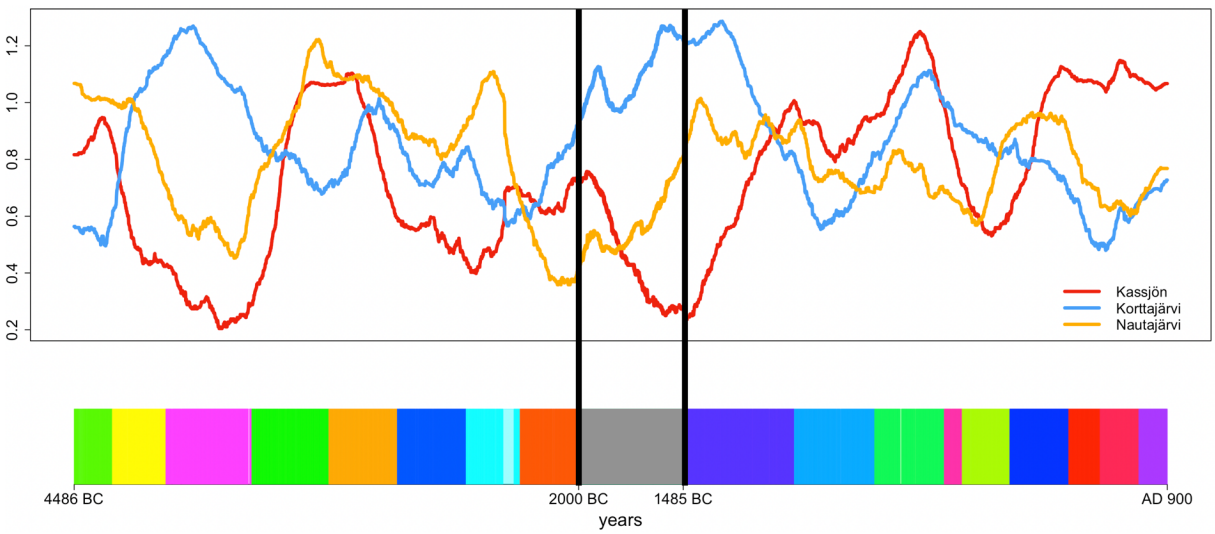


Figure 3.15: Plot of eighteen ($L = 300$) Voronoi cells (bottom) and plot of yearly entropy for each lake (top) when $K = 4$. The horizontal bar below has been obtained by plotting one vertical line for each year in the dataset. The color of each line is determined according to the cell in which a specific year is contained.

Initialization and relabeling

Using the same methodology of the "vanilla" k-means, the climate representatives in the second-level clustering model outlined in Algorithm 3.2 need to be initialized as K randomly chosen cell representatives. To cope with the random initialization, our modified version of the k-means model presented in Algorithm 3.2 is run one hundred times, each starting with a different set of randomly selected cluster templates. The final labels are the result of the aggregation of all parallel runs of the modified k-means. Although each of them is applied to the same cells, these runs are independent, therefore climate cluster $j \in \{1, 2, ..., K\}$ according to a specific run of the proposed k-means algorithm might actually correspond to climate cluster $\tilde{j} \neq j$, according to another run of the same model. To ensure that the clusters from parallel runs of the algorithm coincide, the relabeling framework proposed in Stephens (2000) is employed.

4 | Model choices

In Section 3.1 we outlined the model used to obtain the yearly first-level labels under the mathematical framework of Srivastava et al. (2011) and Tucker et al. (2012) in order to partition the 5271 years into five weather types. Secondly, in Section 3.2, we showed how to cluster the first-level labels into K climate types using the Bagging Voronoi approach to take into account the temporal dependence in the datasets. An entropy-based, lake- and cell-specific weighing procedure has been employed to merge the information contained in the three lakes when labeling each cell. Moreover, the number of bootstrap iterations in the BV algorithm has been set to five hundred, unless specified otherwise. We will now present the results of the second-level clustering, together with the steps followed to tune the hyperparameters involved in the model. From now on, the collection of the first- and second-level clustering will be called Double Clustering Bagging Voronoi (DCBV).

4.1. Tuning the temporal resolution parameter L

The temporal resolution parameter L represents the average number of years contained in each cell and hence directly influences the average number of consecutive years that are clustered to the same climate type. According to the literature (e.g. Secchi et al. (2013), Pozzari (2017), Abramowicz et al. (2019)), L is chosen by fixing K and selecting an unevenly spaced grid of resolution parameters. Then, the second-level clustering for each value of L in the grid is performed and the average yearly entropy $\bar{\eta}$ is stored from Algorithm 3.1. This entails that for a specific choice of K, the following formula is computed:

$$\bar{\eta} = \bar{\eta}(L; K) = \sum_{i=1}^{n} \eta_i / n \tag{4.1}$$

where η_i is the entropy attributed to year i by our model and has been defined in step 21 of Algorithm 3.1. The resolution parameter L is then chosen as a local minimum of $\bar{\eta}(L; K)$. Through this approach, L will be associated with a resolution value that locally minimizes the overall level of uncertainty inherent in the model. The grid for L should be unevenly spaced because the number of nuclei (and thus cells) is dictated by the following

formula:

$$N = \left[\frac{n}{L}\right] = \left[\frac{5271}{L}\right]$$

For example, when L is low (the limiting case being $L = 1$), even minor modifications to the resolution level can lead to substantial fluctuations in N. On the contrary, when L is high (at most $L = n = 5271$), in order for the number of cells to decrease significantly, the resolution should increase considerably.

Figure 4.1 represents the average yearly entropies for $K \in \{3, 4, 5, 6\}$. In order to compare the clustering for different numbers of climate types, the choice for L should be comparable across different values of K. The red points show which values of L lead to the loss of one or more clusters throughout the clustering procedure detailed in Section 3.2.3. When looking for six climate types, at least one cluster is lost for most values of L (see the bottom-right plot in Figure 4.1), therefore from now on we will only explore $K \in \{3, 4, 5\}$. In addition to local minima, another aspect to consider when choosing L is that choosing higher values for the temporal resolution parameter will not only lead to some of the clusters being lost, but it will also imply a less gradual separation between two consecutive climate types. We propose to use $L = 100$ for all number of clusters, based on the local minimum reached around this resolution when $K = 3$ and $K = 5$. Nevertheless, as noted in Secchi et al. (2013), there is no guarantee that tuning L using the average yearly entropy is the optimal or most scientifically rigorous approach. Indeed, given that the temporal resolution parameter has a direct impact on the average number of consecutive years which are assigned to the same climate type, the opinion of experts in this field might help in tuning L.

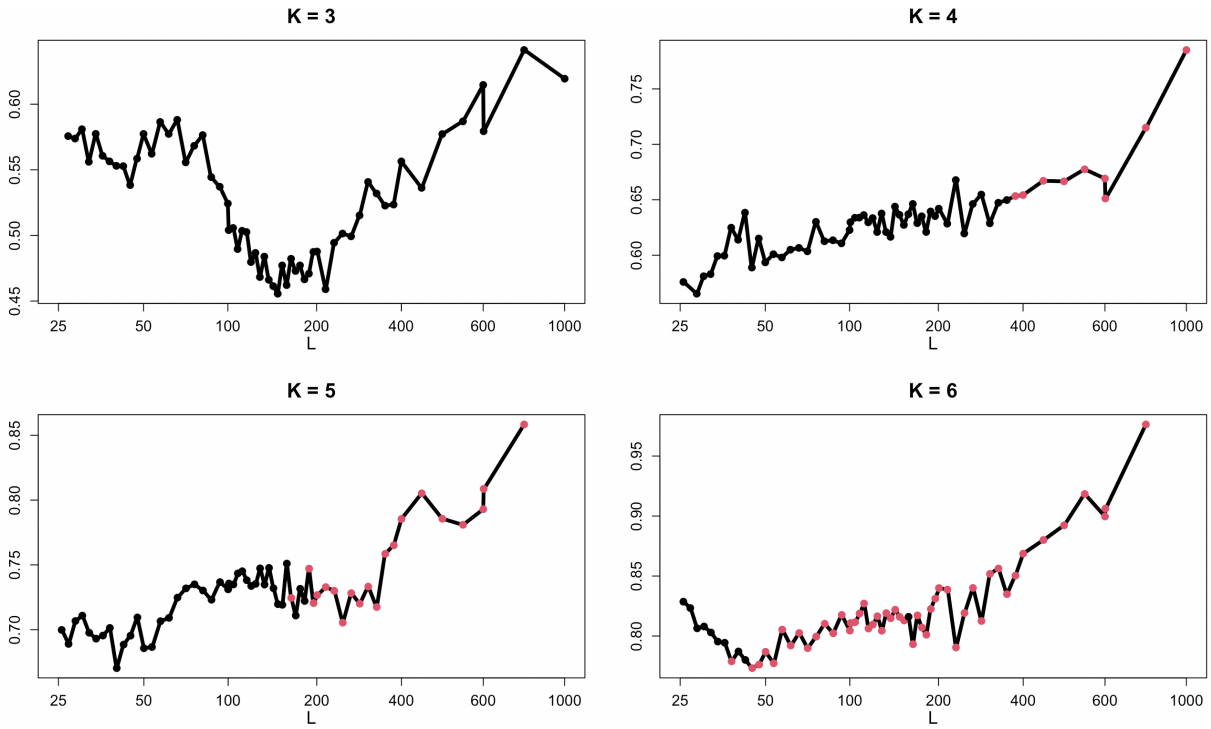


Figure 4.1: Entropy plot for $K \in \{3, 4, 5, 6\}$ as a function of L. The red dots correspond to the values of temporal resolution with respect to which one or more climate clusters are lost throughout the iterations of the model.

4.2. Selecting the number of climate types K

To choose the number of climate clusters K, the same technique detailed in Section 3.1.3 is used: parameter K is increased as much as possible and as long as the addition of the new climate type creates a relevant partition of the years. Contrary to the first-level clustering, we will tune K by inspecting the climate labels directly, rather than the shape of each climate cluster representative. We will do so by using Figure 4.2 as reference: indeed, this plot shows three horizontal bars, corresponding to the clustering of the years with different values of $K \in \{3, 4, 5\}$. Each bar represents 5271 adjacent vertical lines, colored according to the climate label assigned to each year in the dataset. When moving from $K = 3$ (bottom) to $K = 4$ (middle), the new red cluster is obtained as a result of the split of the purple cluster that existed when $K = 3$. Finally, the new green cluster formed when $K = 5$ (top) covers the most recent centuries, labeled as belonging to the orange and gray clusters when K is four. The climate labels for $K = 6$ are not presented in this plot, since for most values of L the clusters found by our model were in reality less than six (see Section 4.1). Since none of the jumps in the values of K seems to contain latent clusters, we decide to set the number of climate types to five. With this choice for

K and L, out of the 5271 years contained in the dataset, 373 (7.1%) are assigned to the gray climate type, 1734 (32.9%) to the orange, 1241 (23.5%) to the purple, 1100 (20.9%) to the red and 823 (15.6%) to the green.

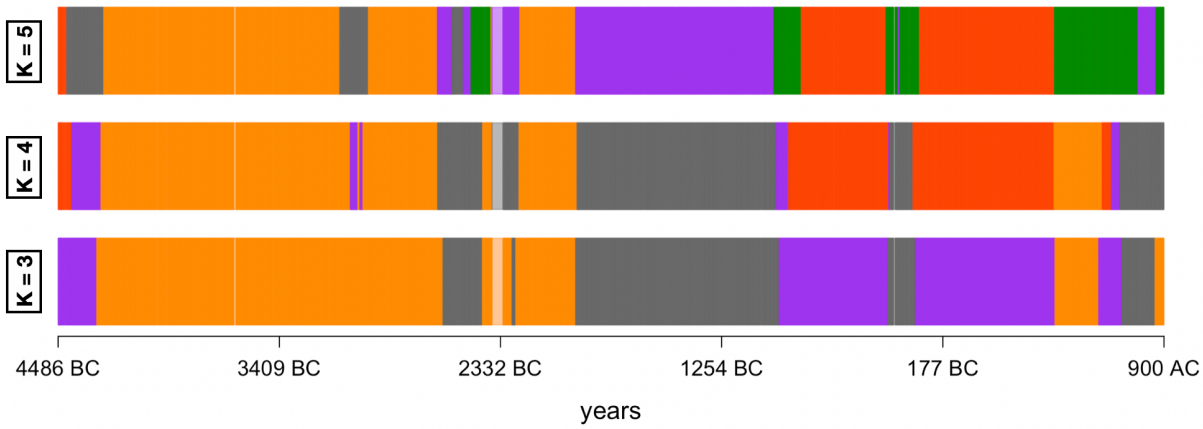


Figure 4.2: Plot of yearly climate labels as a function of K, from three (bottom) to five (top)

We will now further analyze the five climate types computed choosing $L = 100$. For each bootstrap replica of Algorithm 3.1 and for each parallel run of the proposed version of the k-means method, the number of lakes that agree with each cell assignment is stored. Using this information, it is possible to classify each cell assignment into a first, second or third type, as detailed in Section 3.2.3 and exemplified in Figure 3.14. When $K = 5$ and $L = 100$, on average 29.2% of cells receive a label proposed by only one out of three lakes (assignment of the first type), in 46.4% of all cell assignments two lakes simultaneously propose the same climate type (assignment of the second type) and in the remaining 24.5% of cases all lakes vote for the same candidate label (assignment of the third type). This means that in the vast majority of cases at least two lakes agree on the climate label to attribute to a specific cell. Moreover, out of all the assignments of the first type, 44.1% are due to Kassjön, 28.3% to Korttajärvi and 27.6% to Nautajärvi. These considerations point out that the lakes "agree" on the climate labels resulting from the choice of $K = 5$ and $L = 100$.

4.3. Analysis on the agreement

As mentioned in the Introduction, one of the main contributions proposed in this work concerns the uncovering of a common climatological pattern shared by all three lakes. Thus, it is necessary to quantify the agreement of each lake with the final climate labels. To achieve this, we propose a model that "adapts" to each lake the second-level labels found through Algorithm 3.1, by comparing them to each lake's weather clusters. We will

then evaluate each sequence of corrected climate labels with respect to the shared ones. To accomplish this, we devised a classification model: given as input a specific lake's weather labels and the joint second-level clustering, the output of this model will be a set of n climatic labels that result when comparing the local information of each lake and the global signal contained in the result of the second-level clustering. We named this model "Supervised Bagging Voronoi" (SupBV in short), since it solves a supervised classification task and it needs to be built upon the Bagging Voronoi model (see Algorithm 3.1). We stress that SupBV is applied to *one lake at a time*, hence we will obtain three sequences of "adjusted" climate types, each of them being the result of the comparison between the joint climate labels and each lake's specific weather pattern.
When building a model based on a Bagging Voronoi structure, it is necessary to represent each Voronoi tessellation's cell and then cluster (in this case classify) the dataset consisting of all cell representatives. Since SupBV is applied one lake at a time, a cell representative is constructed using the univariate version of the three-dimensional cell representation procedure explained in Section 3.2.2: the frequency in which the weather labels of a specific lake occur within a cell is used to represent the cell itself. The classification model $\mathcal{M}$ consists in employing the Jensen-Shannon divergence (3.4) to assign each cell's representative to its closest climate cluster template, selected from the K cluster representatives identified by the second-level clustering procedure. In short, each cell receives a label according to which is the most similar climate cluster template found according to DCBV. This is why the result of SupBV can be interpreted as a "tailored" or "adjusted" version of the joint climate that matches each lake's specific weather sequence. Figure 4.3 shows the result of SupBV applied to each lake when $K = 5$ and $L = 100$. The colors in the plot match the ones obtained through our second-level clustering model, the result of which is shown in the bottom part of Figure 4.3. From a visual inspection, it is clear that the corrected climate labels share many similarities with the result of the joint clustering. For example, all lakes agree on the red cluster attributed to the years around 177 BC and the orange-colored time frame prior to 3409 BC. Furthermore, both Kassjön and Korttajärvi assign the most recent time period to the green climate type and the purple one appears roughly in the same time frame for Kassjön and Nautajärvi.

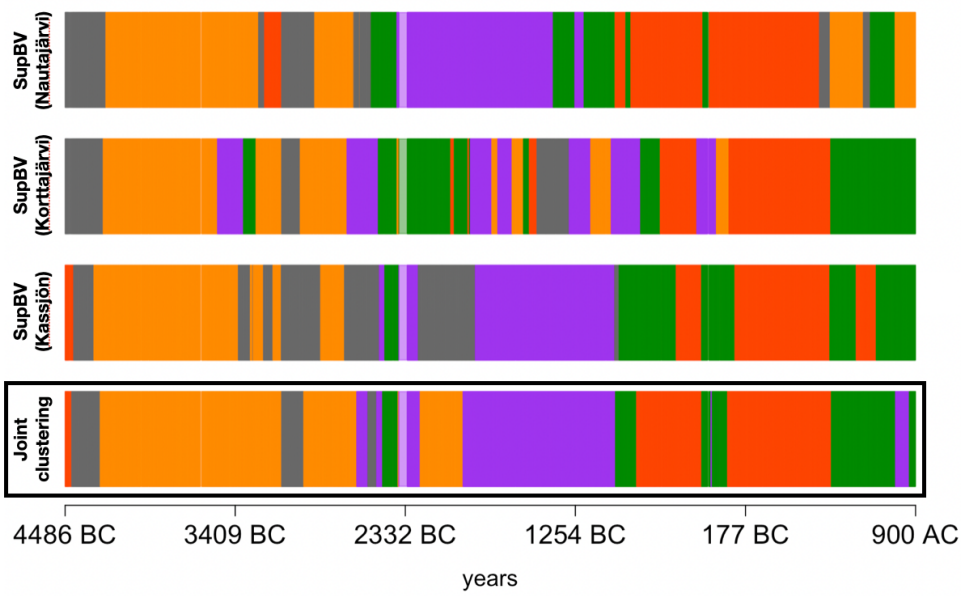
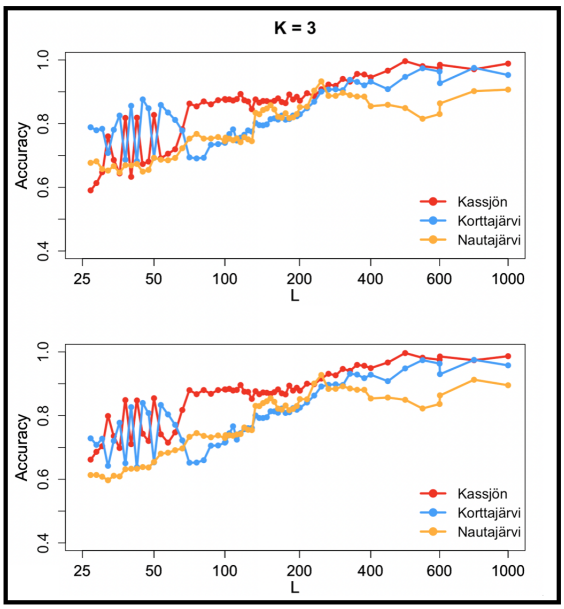


Figure 4.3: The top three horizontal bars show the output of SupBV applied to each lake. The fourth plot enclosed in the black box is the result of our clustering model when $K = 5$ and $L = 100$ and serves as a reference for the top three.

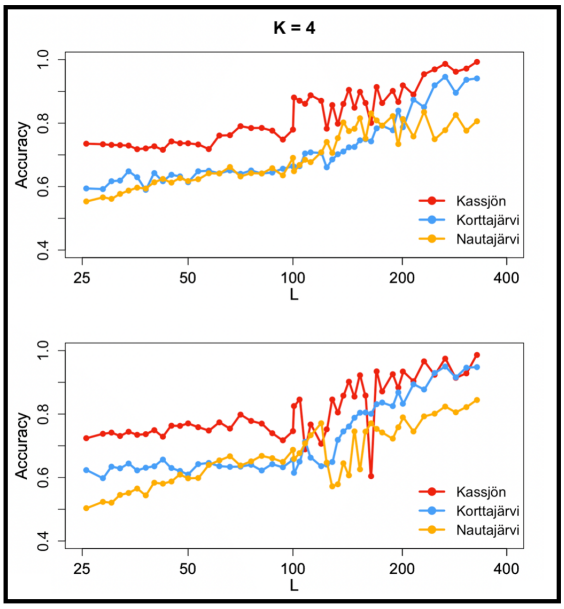
To quantify the agreement of each sequence of adjusted climate labels to the one obtained by jointly clustering the years, we use as classification metric both accuracy and recall. In particular, accuracy is computed as the proportion of years that are assigned to the same climate cluster both by SupBV and by the second-level clustering model. Since the green and the gray climate types have a lower cardinality, it was deemed appropriate to track the performance of our model (in terms of agreement) using the recall as an additional metric. Ideally, if all lakes perfectly agreed with the shared climate labels, then the output of SupBV would match precisely the result of the joint second-level clustering, leading to values of accuracy and recall equal to 100% for all lakes. However, this scenario is extremely unrealistic, given the significant and inherent differences between the lakes in terms of geographical distance, sedimentation procedure and additionally in terms of the dissimilar interpretation of the weather types. Furthermore, the goal of this thesis is not just to produce a single sequence of (climate) labels with which all lakes *perfectly* agree: indeed, the reconstructed clusters should also uncover the common climate shared by the lakes. In other words, the trade-off between appropriate climate reconstruction and high lakes' agreement should not excessively favour the latter.

Figure 4.4 displays the values of accuracy and recall for all values of L and for $K \in \{3, 4, 5\}$, proving that in all lakes such scores are mostly higher than 60%. Table 4.1 shows that, when applying SupBV to the result of our model when $K = 5$ and $L = 100$, all lakes reach high values of accuracy and recall, coherently with the visual inspection in Figure

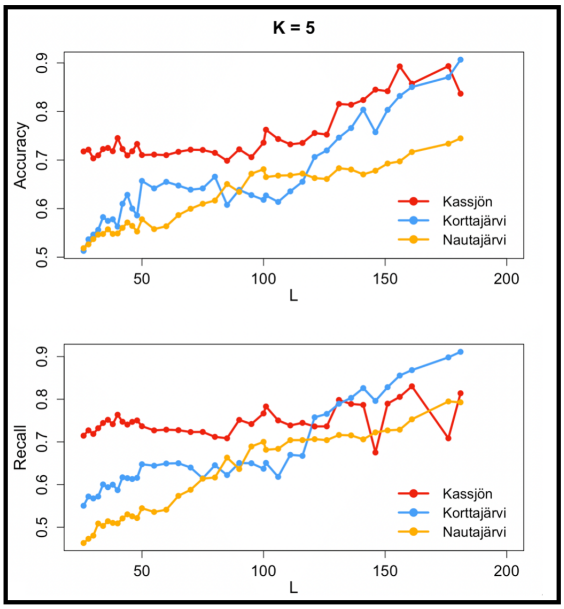
4.3. In order to display the robustness and consistency of these classification scores, we ran SupBV one hundred times and computed the average values of accuracy and recall (see Figure 4.5).



(a)



(b)



(c)

Figure 4.4: Accuracy and recall for all L and $K \in \{3, 4, 5\}$. The x-axis contains only the values of L which did not lead to the loss of clusters.

Lake	Accuracy	Recall
Korttajärvi	73.4%	76.2%
Korttajärvi	63.2%	65.6%
Nautajärvi	67.1%	68.9%

Table 4.1: Accuracy and recall when $K = 5$ and $L = 100$ for all lakes.

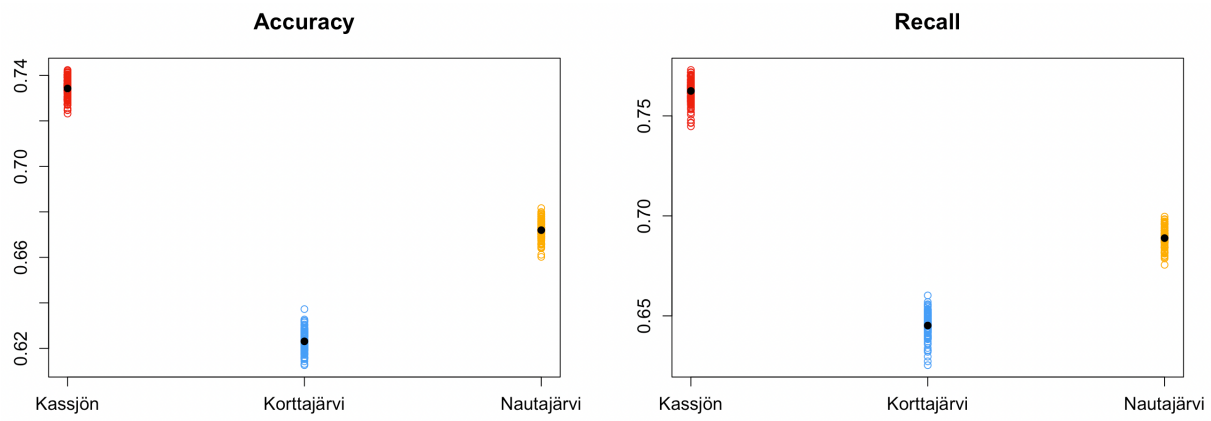


Figure 4.5: Plot of the results from one hundred simulations of SupBV applied to the second-level labels when $K = 5$ and $L = 100$ to show the consistency in the agreement. The black dots are the average accuracy and recall, which are reported in the text. The standard deviations across the one hundred simulations are around 0.004 for the accuracy and 0.006 for the recall.

To assess the efficacy of our results about each lake's agreement with the climate labels, we constructed a naive model, called RandomBV, which entirely disregards the common information among the three lakes when assigning the second-level labels. Indeed, instead of using the iterative procedure detailed in Section 3.2 to cluster the cells in each Voronoi tessellation, RandomBV assigns each cell a *random* label. Then, using SupBV applied to the output of RandomBV, we collect the agreement scores for all lakes. By running this routine ten times, the agreement scores reached by this naive model are as low as 33% (accuracy) and 25% (recall), less than half of our results. The purpose of this analysis is solely to demonstrate that utilizing the Bagging Voronoi method to address the temporal structure of the 5271 years does not necessarily lead to a high degree of agreement in terms of accuracy and recall. Finally, we remark that the classification scores used to quantify each lake's agreement with the result of the joint second-level clustering are never actively used as optimization metrics throughout the algorithm outlined in Section 3.2.3 and are therefore a sole byproduct of the steps implemented in our model to merge the shared information across the lakes.

The analysis and interpretation of each climate cluster is a crucial step to reconstruct the climatic changes pertaining to the Fennoscandian Peninsula from 4486 BC to 900 AC. Such analysis is detailed in Chapter 5.

4.4. Stretching the timeline - from unsupervised to supervised

Having fixed K to five and L to one hundred, each year has now been assigned to one out of five climate types. As pointed out in Chapter 2, the timeline shared by all lakes spans from 4486 BC to AD 900, however Figure 2.8 illustrates that each lake provides data from a wider range. For example Kassjön includes years from AD 900 to roughly AD 2000 and the varves from the Finnish lakes reach as far as 8000 BC. We acknowledge that the scope of our results is constrained by the temporal boundaries within which our model has identified the climate clusters, namely the timeline shared by all lakes. However, it is reasonable to assume that the climate types found by our model are valid in a *neighborhood* of the time frame in which they have been found. In this section, we describe how to use the result of our unsupervised clustering model to shed light on the climate of years beyond the shared timeline 4486 BC - AD 900.

Consider the addition of one thousand new functions concerning the time interval from AD 901 to AD 1900 to lake Kassjön's dataset. In order to understand the climate in these newly added years, it is first necessary to obtain the first-level (weather) labels for each of them. To achieve this, each of the one thousand newly included functions is aligned to each of the five weather clusters in Kassjön, which are shown in the top row of Figure 3.12. Every newly added year is then assigned to a weather cluster, according to which is the closest weather template after the alignment procedure. The framework used to align the new functions to each weather template is the one presented in Srivastava et al. (2011) and briefly described in Section 1.4.2. The new dataset consisting of 6271 weather labels is given as input to SupBV, together with the climate representatives which have been found through our unsupervised clustering model. Figure 4.6 shows that the one thousand years which have been added to the dataset of lake Kassjön are mostly assigned to the green climate type. This is coherent with our model, which pointed out that the most recent time period has been assigned to the green climate.

The same analysis has been conducted on the Finnish lakes, however the timeline was extended to the left, i.e. by adding one thousand years from 5486 BC to 4487 BC. Figure 4.7 shows that according to both Korttajärvi and Nautajärvi the varves immediately preceding 4487 BC are clustered to the gray climate type. This climate cluster is precisely

the one that our model assigns to the earliest segment of the shared timeline.

These results show that the climate clusters found with our model when $K = 5$ and $L = 100$ are still valid in a neighborhood of the timeline shared by all lakes. Nevertheless, investigating the extent of this neighborhood remains a challenge. Using the results found through DCBV to reconstruct more recent climate changes is non appropriate, since it is unreasonable to assume that the same climate types which have been captured by our model are still valid much further than AD 900. Furthermore, after the first half of the nineteenth century human activity has significantly impacted the varves and the sedimentation procedures in the three lakes (see Arnqvist et al. (2016)).

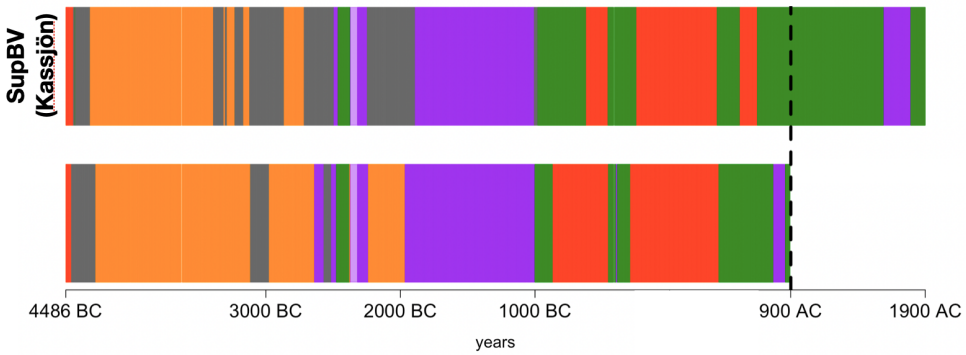


Figure 4.6: The figure at the bottom shows the second-level climate labels obtained when $K = 5$ and $L = 100$. The horizontal bar on the top shows the result of SupBV applied to the extended timeline for lake Kassjön. The dashed black line shows the most recent year in the shared timeline (i.e. AD 900).

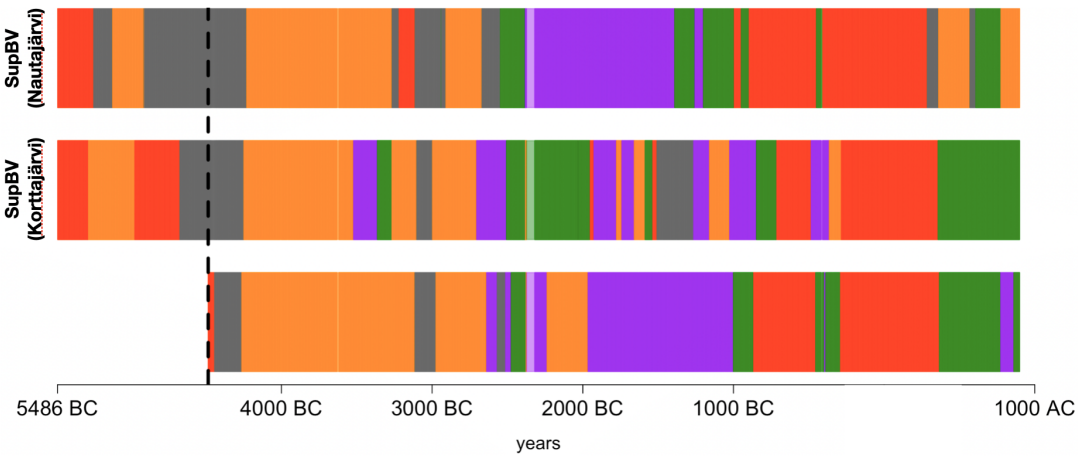


Figure 4.7: The horizontal bar at the bottom shows the second-level climate labels obtained when $K = 5$ and $L = 100$. The bars at the top and in the middle show the results of SupBV applied to the extended timeline for Nautajärvi and Korttajärvi. The dashed black line shows the earliest year in the shared timeline (i.e. 4486 BC).

5 | Climate interpretation

In Chapter 3 the methodology followed to cluster the years into K climatic types has been presented. In Chapter 4 we showed how to tune the parameters of DCBV and proved that the three lakes agree regarding the second-level climate labels. Throughout this Chapter, we will detail how to interpret each climate type obtained with $K = 5$ and $L = 100$.

5.1. Climate clusters interpretation

As explained in Section 3.2.2, each climate cluster representative is a collection of three distributions, describing the frequency of occurrence of each weather type inside the climate cluster itself. Figure 5.1 shows such frequencies for each climate cluster when $K = 5$ and $L = 100$. Based on Section 3.1, the weather types of each lake differ from the ones of the other two. Moreover, Figure 5.1 shows that it is not straightforward to "match" the climate cluster representatives across the lakes: indeed, each climate cluster is identified by a lake-specific mixture of weather types. For example, none of the climate clusters representatives in Korttajärvi displays a similar weather pattern as the one of the first climate type CL1 in Kassjön. The difficulty in matching different climates across the lakes is increased by the fact that, as pointed out in Section 3.1, it is not straightforward to find similarities across the first-level (weather) cluster representatives of different lakes. These considerations entail that, even though each lake agrees with the partition of the years induced by the second-level clustering (as proven in Section 4.3), climate is expressed uniquely by each lake. In other words, even though all lakes share the same climate clusters across the 5271 years, the interpretation of each climate type in terms of weather patterns is specific to each lake. To thoroughly show and explain this concept, it is essential to conduct a chronological examination of the climatological evolution of each lake, using Figure 5.2 as a reference. The time frame concerning the data spans from 4486 BC to AD 900: the period from 4486 BC to 4267 BC is characterized by a gray-colored climate, which is then followed by the orange-colored one from 4267 BC to 2640 BC. The years ranging from 2640 BC to 1965 BC seem to be clustered to different climate types: indeed, this time frame could represent a transition from the long-lasting orange cluster

to the almost equally long purple one. The latter lasts from 1965 BC to 1000 BC and is followed by an alternation of the red and green clusters, until AD 350. Finally, the most recent years in the timeline (from AD 350 to AD 900) are clustered to the green climate type.

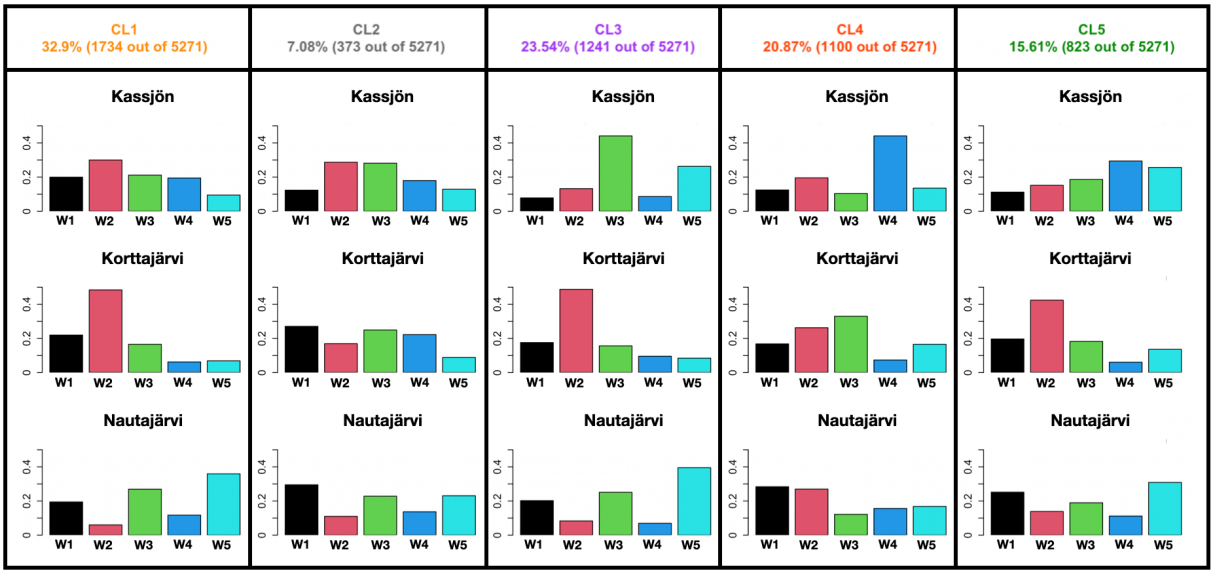


Figure 5.1: Each column shows a single climate cluster representative. The color of the histograms is the same across the lakes only for simplicity: weather Wj does not have the same interpretation for Kassjön, Korttajärvi or Nautajärvi (see Section 3.1).

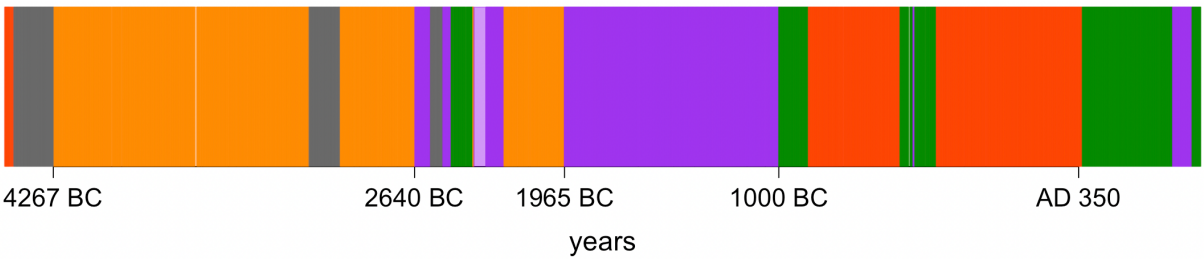


Figure 5.2: Yearly climate labels ($K = 5$, $L = 100$)

5.1.1. Climate clusters for Kassjön

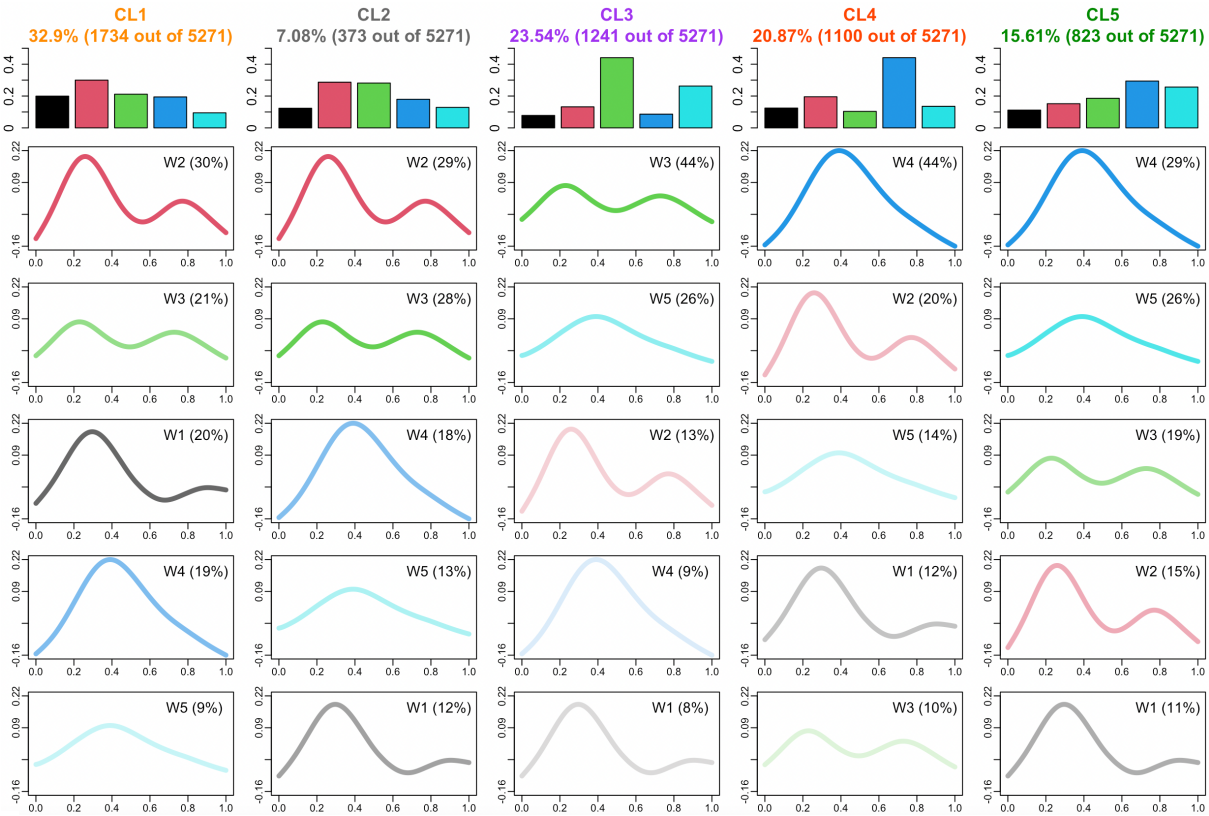


Figure 5.3: Climate cluster representatives for Kassjön. The first row contains the histograms representing each climate according to lake Kassjön. Each column shows the weather representatives for Kassjön ordered and shaded from most (top) to least (bottom) frequently appearing in each climate cluster. The percentage in the top-right corner of each plot shows how many times the weather type displayed in the box appears in the climate cluster. Consider for example the column associated to CL1: the most frequently appearing weather label is the red one, followed by the green, then black and finally the blue and cyan.

Based on Figure 5.3, the relatively short time period **from 4486 BC to 4267 BC** in lake Kassjön assigned to the gray climate cluster (CL2) is characterized mainly by a "bimodal" pattern (W1 and W3), with a particularly intense spring peak followed by another one in late summer. The orange cluster (CL1) spanning **from 4267 BC to 2640 BC** can be interpreted similarly to the gray one, the main difference is an inversion in the frequency of appearance of weather clusters W1, W4 and W5. The time frame **from 2640 BC to 1965 BC** can be interpreted as a transition from the long lasting orange cluster to the purple one. During this period the appearance of the purple and green clusters may indicate the start of a transition from a bimodal weather pattern to an "unimodal" one, characterized by having only one peak. After 2640 BC, the bimodal patterns start to flatten, as their spring peaks become lower and lower: during the purple

climate cluster (CL3) lasting **from 1965 BC to 1000 BC** most years have very mild spring and autumn peaks (W3). The unimodal weather tendencies are starting to become more frequent (W5). The time period **from 1000 BC to AD 350** is characterized by an alternation of the red (CL4) and green (CL5) climate clusters. The years clustered to these climate types show a pronounced inclination towards a single mode or peak (W4 for both climates and additionally W5 for the green one), completing the transition started around year 1965 BC. Notice that the mild bimodal weather patters in the recently concluded purple cluster are still present in the red climate (in the form of W2) as the second most frequently occurring weather type. Finally, the timeline ends with a prevalence of unimodal weather tendencies (W4 and W5) in the green climate cluster (CL5), which occupies the last period **from AD 350 to AD 900**.

5.1.2. Climate clusters for Korttajärvi

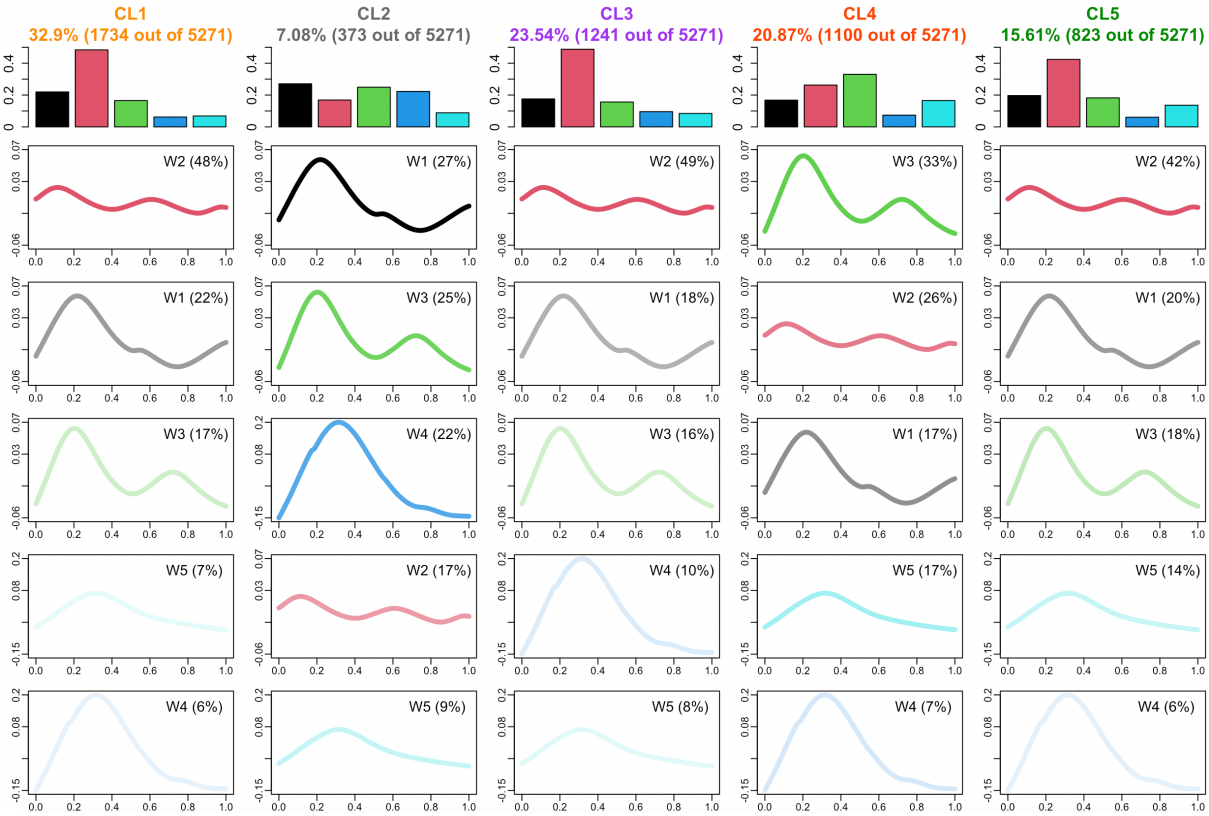


Figure 5.4: Climate cluster representatives for Korttajärvi. The first row contains the histograms representing each climate according to lake Korttajärvi. Each column shows the weather representatives for Korttajärvi ordered and shaded from most (top) to least (bottom) frequently appearing in each climate cluster. The percentage in the top-right corner of each plot shows how many times the weather type displayed in the box appears in the climate.

Similarly to the interpretation of the initial gray cluster (CL2) for Kassjön, in lake Korttajärvi the time frame **from 4486 BC to 4267 BC** is characterized mainly by unimodal weather patterns, with possibly a very low peak during late spring, as suggested by W1 and W3. When compared with the others, the gray climate type is the only one containing a considerable amount of years assigned to weather type W4. The orange period **from 4267 BC to 2640 BC** (CL1) shares some similarities with the gray one: indeed, the most frequently appearing weather clusters in the gray climate (W1) has now moved from the first to the second place and has instead been replaced by W2. Such weather pattern is characterized by having extremely moderate spring and autumn peaks. As indicated before, the period spanning **from 2640 BC to 1965 BC** presents many "conflicting" climatic behaviours, possibly because of its position between two long lasting climate types. Most climates in this time frame share the predominant occurrence of weather cluster W2. Korttajärvi's purple cluster (CL3) **from 1965 BC to 1000 BC** shares many similarities with the orange one, as it contains years that don't seem to experience high inflows of either minerogenic or organic material (W2). Notice that the flat behaviour of W2 in Korttajärvi greatly resembles the shape of Kassjön's weather cluster W3. For this reason, since in the purple climate type of Kassjön contains mainly years assigned to weather type W3 and Korttajärvi to W2, the interpretation of the purple climate cluster (CL3) according to Korttajärvi is quite similar to the one expressed by lake Kassjön. The alternation of the red (CL4) and green (CL5) climatic clusters **from 1000 BC to AD 350** reflects a fluctuation between weather patterns with and without a spring peak: indeed, the most frequently appearing weather type is W2 for the green climate and W3 for the red one. Notice that the transitions between these two climate types do not appear sudden or abrupt, since W2 is both the first and the second most frequently appearing weather type in the green and the red climate types, respectively. Moreover, the second most frequently appearing weather type in CL5 (which is W1) presents some similarities with the first most frequently appearing one in CL4 (i.e. W3): indeed, both weathers show a high spring peak. Finally, the green colored years **from AD 350 to AD 900** (CL5) are characterized mostly by moderate, almost flat weather patters.

5.1.3. Climate clusters for Nautajärvi

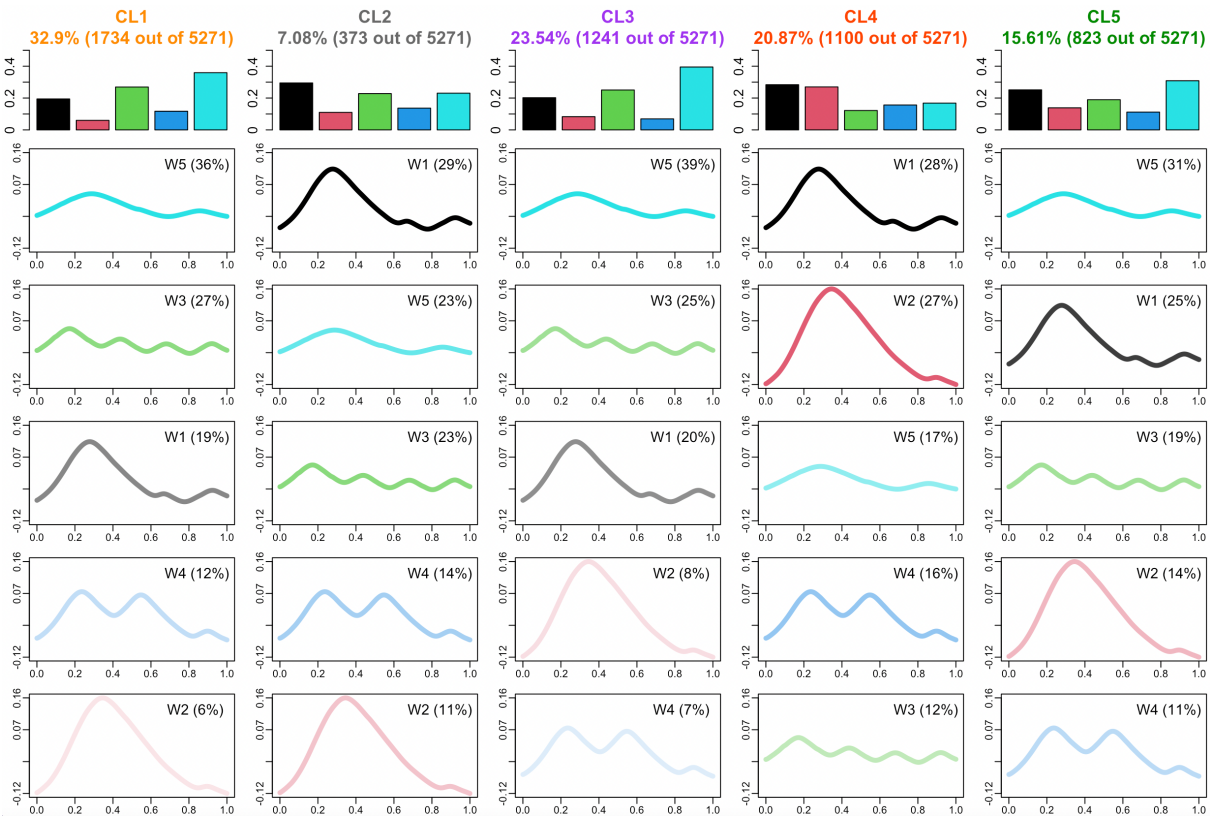


Figure 5.5: Climate cluster representatives for Nautajärvi. The first row contains the histograms representing each climate according to lake Nautajärvi. Each column shows the weather representatives for Nautajärvi ordered and shaded from most (top) to least (bottom) frequently appearing in each climate cluster. The percentage in the top-right corner of each plot shows how many times the weather type displayed in the box appears in the climate.

The gray climate (CL2) appearing **from 4486 BC to 4267 BC** in Nautajärvi primarily contains years whose weather has a prominent spring peak, followed by a slight increase during winter (W1). The second most frequent weather type in the gray cluster becomes the first most prevalent one in the orange colored time frame **from 4267 BC to 2640 BC** (CL1). Roughly 27% of the years contained in CL1 are clustered to the third weather type (W3), which appears to have many moderate local peaks. Overall, CL1 seems to contain years in which there hasn't been an imbalance in clastic-organic inflow, which would have lead to more prominent peaks. Notice that the orange cluster is characterized by very mild peaks both in Korttajärvi and in Nautajärvi, contrary to Kassjön. As already suggested previously, the period **from 2640 BC to 1965 BC** is quite chaotic. Upon closer inspection of the clusters appearing in this time frame, CL1, CL3 and CL5 can be interpreted in similar ways according to the Finnish lakes in terms of shape of the most

frequently appearing weather cluster. Exactly as in Korttajärvi, the purple period (CL3) is quite similar to the orange one: the weather of most years **from 1965 BC to 1000 BC** has extremely mild spring peaks (W5), followed by an almost equal mixture of weather types W1 and W3. The switch from the purple (CL3) to the green (CL5) climate around 1000 BC is quite gradual, since both clusters share the same most frequent weather type (W5). However, there is a clear difference between CL3 and CL5: indeed, the latter contains more years with a more noticeable spring peak (W1). For what concerns the period **from 1000 BC to AD 350**, while CL5 is mainly characterized by mild weather patterns, the red cluster CL4 clearly depicts unimodal weather patterns with a modest peak later in the year (W1 and W2). Note that the alternation between the red and green climates shows affinities with the same time frame in lake Kassjön. Finally, the last years **from AD 350 to AD 900** are assigned to the green climate cluster CL5, which, according to Nautajärvi, mostly contains weather types with a moderate spring peak.

5.1.4. Final remarks

As pointed out in these sections, it is possible to appreciate that the Finnish lakes exhibit greater similarities amongst themselves than with Kassjön. Consider for example the range and overall shape of their weather clusters' representatives and their climatological history detailed throughout these sections. This fact is highlighted in Figure 5.6, in which the content of each climate is shown for each lake in terms of the first and second most frequently appearing weather types in each cluster. Based on the inspection of the content of the climate types conducted in the previous sections, it is possible to conclude that each lake exhibits a unique expression of climate in terms of mixture of weather types, however still preserving a common and shared message. This "common thread" among the lakes consists firstly in the timing of the climate changes (since we have been able to obtain a single sequence of climate labels valid for all lakes) and secondly to some overlap in the content of some climate types. Consider, for example, the similarities across the lakes with respect to the content of climate cluster CL3 in Figure 5.6. Based on this plot, it is possible to highlight the similar shapes of the top two most frequently appearing weather types in CL2. This is particularly striking, considering the quite different shape of the lakes' weather types detailed in Section 3.1. The same, however, cannot be argued for climates CL1 and CL5. Nevertheless, it is not unreasonable that the lakes don't share exactly the same interpretation for each climate, considering the different geographical location of the lakes, the different methodology used to gather the data (see Sections 2.1 and 2.2) and the possibly different seasonal layering procedures which have lead to the creation of the varves in each lake.

Our efforts have helped in proving that climate changes are expressed in different ways by each lake and the uncovered common denominator in the three datasets is precisely the occurrence and timing of such climate changes. Furthermore, an additional notable result from our analysis in Sections 5.1.1, 5.1.2 and 5.1.3 is that the variations in climate happen gradually for all lakes: indeed, weather patterns of consecutive climate types share many similarities.

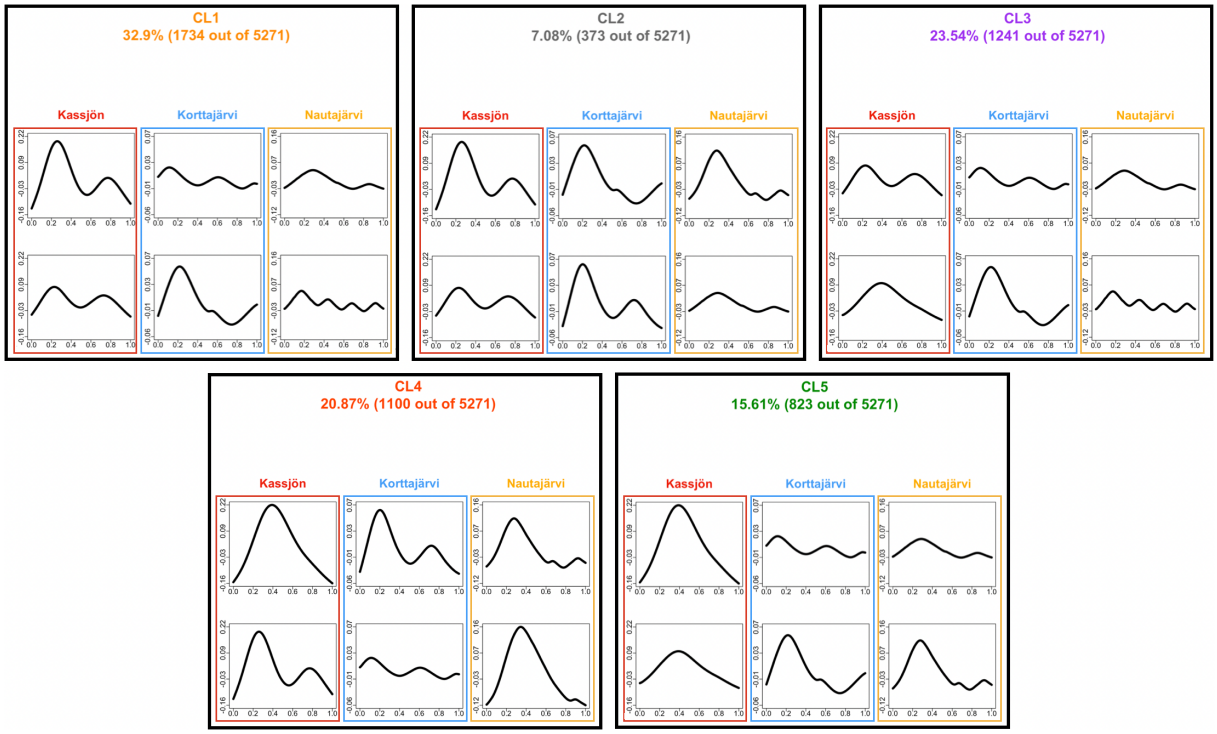


Figure 5.6: Comparison of the five climate clusters across the lakes. Each box represents the content of each climate type, in terms of the first (top) and second (bottom) most frequently appearing weather clusters for each lake.

6 | Conclusions and future developments

The goal of this work has been to cluster temporally-dependent misaligned functional data into robust and homogeneous climate types, merging the information coming from three lakes, in an attempt to reconstruct past climatic changes regarding the Fennoscandian Peninsula.

We started by studying firstly the scientific literature regarding the climatological information contained in the three datasets and the techniques used to collect them, secondly the broad mathematical literature about the task of functional data registration and thirdly we reviewed papers and articles based on the same data and aimed at similar goals. Specifically, the initial portion of this research has been committed to examining the existence of misalignment in the data and determining a suitable mathematical framework to effectively address it. We started by applying the methodology of Sangalli et al. (2010) and then, given the drawbacks of this framework in the context of the specific task at hand, moved to the one proposed by Srivastava et al. (2011) and Tucker et al. (2012). We investigated a family of models rooted in the Bagging Voronoi algorithm to divide the years in the dataset into a suitable number of climate clusters. In this work we presented two of those, namely Bagging Voronoi K-Mean Alignment (BVKMA, briefly described in the Appendix) and Double Clustering Bagging Voronoi (DCBV). The former is used to obtain the climate clusters in one single step and its name comes from the fact that it is an extension of the vanilla k-means algorithm to the task of clustering misaligned and temporally- or spatially-dependent data. DCBV, on the other hand, consists in obtaining the climate labels by firstly clustering the weather in each lake and then using these three sequences of labels to reconstruct the shared climatic history of the three lakes. To do so, previous works on the same topic (such as Arnqvist et al. (2016) and Pozzari (2017)) reconstructed the weather labels in each lake using the mathematical framework proposed in Sangalli et al. (2010). In this work the functional clustering model proposed in Srivastava et al. (2011) has been applied instead. Although DCBV has already been

applied in previous papers for a similar purpose (such as Abramowicz et al. (2019)), this thesis represents the first attempt to generalize this model to a "multi-lake" scenario, in which data from multiple sources are simultaneously taken into account. To do this, we devised a technique called *soft-majority vote* and designed a weighing procedure to allow to obtain a single sequence of climates with which all lakes agree. Not only these model choices allow to separate the contributions of each lake to the final clustering, but they also allow some flexibility throughout the second-level labeling process, especially through the weighing procedure. Within the context of this thesis, we worked on data coming from only three lakes, however our model can straightforwardly be applied to the case in which signals collected from additional lakes are added to the joint dataset.

We dedicated the next portion of our work to tuning some model parameters, including the number of weather and climate clusters. To summarize, we propose to use $k = 5$, $K = 5$ and $L = 100$. Moreover, we designed two models to quantify the degree in which each lake agrees with the result of DCBV. These models have been named Supervised Bagging Voronoi (SupBV) and Random Bagging Voronoi (RandomBV). Given the sequence of shared climate labels and a specific lake's weather labels, SupBV is able to correct and adjust the result of DCBV to suit each lake's own weather signal, which is then compared with the result of the second-level clustering. Our model reaches consistently high values of agreement for each lake, according to conventional classification scores (accuracy and recall). We remark that such metrics are never actively used in the DCBV model. An additional benefit of SupBV consists in its application to test the extension of the second-level clusters found by our model to the years in a neighborhood of the timeline spanned by the joint dataset. On the other hand, RandomBV serves as a simple baseline model to compare the double clustering procedure we proposed in terms of agreement. To be more precise, this naive reference model is used to check whether the high values of accuracy and recall reached by our model are solely due to the methodology used to take into consideration the temporal dependence of the signals in the datasets. The results of RandomBV show that it is not guaranteed that using the Bagging Voronoi approach to account for the temporal structure of the data will necessarily lead to high levels of accuracy and recall. Indeed, although RandomBV is based on the structure of the Bagging Voronoi algorithm, this baseline model leads to values of recall and accuracy which are less than half of the ones reached by our proposed model.

Finally, the main contribution of this thesis stands out when analyzing the content of each climate cluster found using DCBV. Through the reconstruction of the climatological history of each lake, we have been able to show that each lake exhibits distinct expressions

of climate change, while the shared characteristic across all three datasets is the occurrence of these climate variations. This result is intuitive: indeed, although they might have been influenced by similar climatological phenomena, the lakes might not share the same varve formation procedures, which is highly influenced by the presence of additional water basins in the vicinity of each lake. Moreover, the geographical distance is another factor contributing to the different expressions of climate. This is coherent with the fact that, in terms of both weather and climate, the Finnish lakes share more similarities among themselves than with Kassjön.

In Chapter 4, we showed that all lakes show agreement scores (i.e. accuracy and recall) higher than 60% for all values of K and spatial resolution parameter L. Based on Figure 4.4, it is clear that the metrics used to quantify the agreement increase as L increases. This is due to the fact that higher values of L lead to more abrupt changes from one climate to the next and thus it will be more likely that the correction of such changes by SupBV will be similar to the result of DCBV. On the other hand, lower values of the temporal resolution parameter are suitable to reconstruct smoother and more gradual climate changes. Moreover, our model is designed to capture both the common message in the three lakes and the temporal structure of climate, hence it is reasonable that the latter influences the accuracy and recall of SupBV with respect to DCBV. For these reasons, possible future developments include a study on how to decouple the effect on the metrics used to quantify the agreement into a contribution of the temporal structure and a contribution of the actual shared message across the lakes. This can help in understanding the proportion of the agreement which is truly due to the goodness of the model in reconstructing the common information of the lakes and which is the proportion which comes from the temporal structure, instead.

In Section 4.3, we pointed out that the high level of agreement of each lake is a byproduct of the model choices implemented in DCBV, since the classification metrics used to quantify such agreement are never actively used in our model. An additional further development consists in implementing a "baseline" model which partitions the years into a suitable number of clusters not based on the shared climate, but focusing only on which is the optimal partition that maximizes the agreement across the three lakes. This can help in understanding the goodness of our results by comparing the agreement scores reached by our version of DCBV with the ones provided by a model whose sole purpose is to find an optimal partition of the years with respect to the agreement scores.

Bibliography

Abramowicz, K., Arnqvist, P., de Luna, S. S., Secchi, P., Vantini, S., and Vitelli, V. (2014). Was it snowing on lake kassjön in january 4486 bc?

Abramowicz, K., Arnqvist, P., Secchi, P., and de Luna, S. S. (2017). Clustering misaligned dependent curves applied to varved lake sediment for climate reconstruction. *Stoch Environ Res Risk Assess*, 31:71–85.

Abramowicz, K., Schelin, L., de Luna, S. S., and Strandberg, J. (2019). Multiresolution clustering of dependent functional data with application to climate reconstruction. *Stat*, 8(1):e240. e240 sta4.240.

Arnqvist, P., Bigler, C., Renberg, I., and de Luna, S. S. (2016). Functional clustering of varved lake sediment to reconstruct past seasonal climate. *Environ Ecol Stat*, 23:513–529.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Guo, X., Wu, W., and Srivastava, A. (2022). Data-driven, soft alignment of functional data using shapes and landmarks.

Ieva, F. and Paganoni, A. M. (2013). Depth measures for multivariate functional data. *Communication in Statistics- Theory and Methods*, 42:1265–1276.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Lopez-Pintado, S., Juan, and Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104:718–734.

Menafoglio, A. and Secchi, P. (2017). Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. *European Journal of Operational Research*, 258(2):401–410.

Ojala, A. and Alenius, T. (2005). 10 000 years of interannual sedimentation recorded in

the lake nautajärvi (finland) clastic-organic varves. *Palaeogeography Palaeoclimatology Palaeoecology*, 219:285–302.

Ojala, A., Heinsalu, A., Kauppila, T., Alenius, T., and Saarnisto, M. (2008). Characterizing changes in the sedimentary environment of a varved lake sediment record in southern central finland around 8000 cal. yr bp. *Journal of Quaternary Science*, 23:765 – 775.

Petterson, G., Odgaard, B. V., and Renberg, I. (1999). Image analysis as a method to quantify sediment components. *Journal of Paleolimnology*, 22:443–455.

Petterson, G., Renberg, I., and et al., P. G. (1993). Spatial uniformity of sediment accumulation in varved lake sediments in northern sweden. *Journal of Paleolimnology*, 9:195–208.

Pozzari, A. (2016/2017). Functional clustering of varved lake sediment to reconstruct past climate in fennoscandia (master thesis). Master's thesis, Politecnico di Milano.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics and Data Analysis*, 54(5):1219–1233.

Secchi, P., Vantini, S., and Vitelli, V. (2013). Bagging voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation*, 22:53–64. Spatial Statistics for Mapping the Environment.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. (2011). Registration of functional data using fisher-rao metric.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Tiljander, M., Ojala, A., Saarinen, T., and Snowball, I. (2002). Documentation of the physical properties of annually laminated (varved) sediments at a sub-annual to decadal resolution for environmental interpretation. *Quaternary International*, 88(1):5–12. The Value of Annually Laminated Sediments in Palaeoenvironment Re constructions: Dedicated to Bjorn E. Berglund.

Tiljander, M., Saarnisto, M., Ojala, A., and Saarinen, T. (2003). A 3000-year palaeoenvironmental record from annually laminated sediment of lake korttajarvi, central finland. *Boreas*, 32:566 – 577.

Tucker, J. D., Wu, W., and Srivastava, A. (2012). Generative models for functional data

using phase and amplitude separation. *Computational Statistics and Data Analysis*, 61(C):50–66.

Vantini, S. (2012). On the definition of phase and amplitude variability in functional data analysis. 21(4):676–696.

A | Appendix

A.1. BVKMA

In this section the description of Bagging Voronoi K-Means Alignment is provided. The family of models under this name is based on an extension of the k-means algorithm to the task of clustering misaligned and temporally (or spatially) dependent functional data. Contrary to DCBV, which is based on a double clustering approach, BVKMA allows to obtain the climate labels without passing through the reconstruction of the weather for each lake. The steps employed in this model follow the ones described in Algorithm 3.1, with some modifications:

- While the DCBV applied the Bagging Voronoi method to the dataset composed of the three weather labels for each year, BVKMA applies BV directly to the functional dataset.
- In the BVKMA algorithm, before computing the cell representative $\bar{\boldsymbol{f}}_c^b$ for each cell c and bootstrap iteration b, the functions belonging to each cell are aligned one lake at a time, according to the framework presented in Section 1.4.2 and in Srivastava et al. (2011).
- For each bootstrap iteration $b = 1, 2, ..., B$, the representative $\bar{\boldsymbol{f}}_c^b$ of the c-th cell with $c = 1, 2, ..., N$ is not a triplet of distributions, but a triplet of functions: i.e. $\bar{\boldsymbol{f}}_c^b = [\bar{f}_c^{1,b}, \bar{f}_c^{2,b}, \bar{f}_c^{3,b}]$. Each component is the pointwise average of the aligned functions belonging to a specific cell and to a specific lake. In particular, defining $N_{c,b}$ as the number of years in V_c^b (i.e. the subset of years within the c-th cell at the b-th bootstrap iteration),
$$\bar{f}_c^{\ell,b} = \frac{\sum_{i \in V_c^b} f_i^\ell}{N_{c,b}} \quad \forall \ell = 1, 2, 3 \tag{A.1}$$
- Model $\mathcal{M}$ used to cluster the three-dimensional cell representatives is based on Algorithm 1.1, extended to the case in which the functions to cluster are multidimensional.

Algorithm A.1 shows in detail the steps followed in BVKMA.

Algorithm A.1 BVKMA

Let $\{\boldsymbol{f}_i\}_{i=1}^n$ be the triplets of functions on the lattice of $n = 5271$ years
Fix number of clusters K, temporal resolution parameter L, number of bootstrap iterations B
Set $S_0 = \{-4486, -4485, ..., 900\}$ as the one-dimensional lattice of years
Define the number of nuclei $N = \left[\frac{n}{L}\right]$
Bootstrap step:
for $b = 1, 2, ..., B$ **do**
  Randomly generate a set of N nuclei $\Phi_N^b = \{Z_1^b, Z_2^b, ..., Z_N^b\}$ such that $Z_c^b \overset{i.i.d.}{\sim} \mathcal{U}(S_0)\ \forall c = 1, 2, ..., N$
  Obtain a random Voronoi tessellation of S_0 in N cells, called $\{V_c^b := V(Z_c^b|\Phi_N^b)\}_{c=1}^N$, by assigning each year in S_0 to the nearest nucleus Z_c^b according to the L_1 or L_2 distance
  for $c = 1, 2, ..., N$ **do**
    for $\ell = 1, 2, 3$ **do**
      Align $\{f_i^\ell\}_{i \in V_c^b}$
    end for
    Compute $\bar{\boldsymbol{f}}_c^b$, a *local* representative of the data in the c-th cell (formula (A.1))
  end for
  Using a suitable model $\mathcal{M}$, cluster the N cell representatives
  for $i = 1, 2, ..., 5271$ **do**
    Assign to the i-th year the label of its cell
  end for
  Return the labels of the b-th bootstrap iteration, which create a partition of the n years into K clusters $\{C_j^b\}_{j=1}^K$
end for
Aggregation step:
Using the framework proposed in Stephens (2000), relabel the cluster assignments $\{\{C_j^b\}_{j=1}^K\}_{b=1}^B$
for $i = 1, 2, ..., 5271$ **do**
  Compute $\pi_i^j = \#\{b \in \{1, 2, ..., B\} : i \in C_j^b\}/B \quad \forall j = 1, 2, ..., K$
  Assign to year i the label $j^* = \underset{j=1,2,...,K}{\operatorname{argmax}}\ \pi_i^j$, i.e. its most frequently assigned label throughout the B bootstrap iterations
  Compute the *entropy* of the i-th year as $\eta_i = -\left(\sum_{j=1}^K \pi_i^j \log \pi_i^j\right)$
end for

Together with the cell clustering model $\mathcal{M}$, steps 1, 9 and 12 of the BVKMA algorithm A.1 represent the new features used to extend the Bagging Voronoi framework to the case of multidimensional functional data. In order to cluster the cell representatives, model $\mathcal{M}$ has been designed as an extension to the "multi-lake" context of the k-means alignment algorithm proposed in Srivastava et al. (2011) and detailed in Algorithm 1.1. Since each cell (as well as each cluster) is represented through a three-dimensional functional object, we use the same approach detailed in Section 3.2.3: to measure how close a cell is to a cluster we propose to use a weighted distance of the components corresponding to the different lakes (see equation (3.3)). The weights are obtained as in Section 3.2.3. According to the mathematical framework described in Srivastava et al. (2011) and briefly summarized in Section 1.4.2, the metric used is the L_2 distance between the SRVF representatives of the functions. In particular, we use the same methodology named *soft-majority vote*: before assigning a cell to a climate type, a preliminary clustering is performed to compute each lake's candidate label. The final label assigned to the cell is found precisely as in equation (3.7). This extension of Algorithm 1.1 has been implemented based on the `fdasrvf` R package and it can be readily applied to the broad task of clustering temporally (or spatially) dependent, misaligned and *multi-dimensional* functional data, as long as suitable ad hoc weights are provided. Note that since BVKMA is applied to a functional dataset constructed with signals sampled in two hundred time points, the computational cost of this model is much higher than the one of DCBV, for the following reasons:

1. Model $\mathcal{M}$ needs to be applied B times.
2. Since a single run of model $\mathcal{M}$ is computationally more expensive than a single run of the second-level clustering in DCBV, a multistart approach cannot be applied: hence, it is necessary to properly choose the initial templates by applying preliminary clustering models to the functional dataset (similarly to the one detailed in Section 3.1.1). Considering that each cluster representative is a triplet of functions, this initialization procedure has to be run three times (one for each lake's dataset).
3. Before applying $\mathcal{M}$ to cluster the three-dimensional cell representatives, it is necessary to align the functions within the cell itself for each of the three lakes.

The computational cost represents one of the main limitations of BVKMA, even though the implementation of the code for this model is based on parallel computing. For example, it is extremely difficult to properly tune the temporal resolution parameter L and the number of climate clusters K. For this reason, we will show the results of BVKMA for $K = 5$ and $L = 100$, which are the values proposed for the parameters according to DCBV. Note that there is no guarantee that the pair (K, L) explored in this section is

the "optimal" one for BVKMA. The following results have been obtained thanks to the computational resources of Kebnekaise, a supercomputer offered by the High Performance Computing Center North (HPC2N).
Figure A.1 shows the yearly climate labels according to one hundred bootstrap iterations of BVKMA when $K = 5$ and $L = 100$. With this choice, one cluster is lost, leaving only four climate types. As previously stated, this result is not worrying, as the couple (K, L) explored in this section is not necessarily the best one for BVKMA. Despite this, as Figure A.1 shows, the timing of the climate changes is quite coherent with the one provided by DCBV.

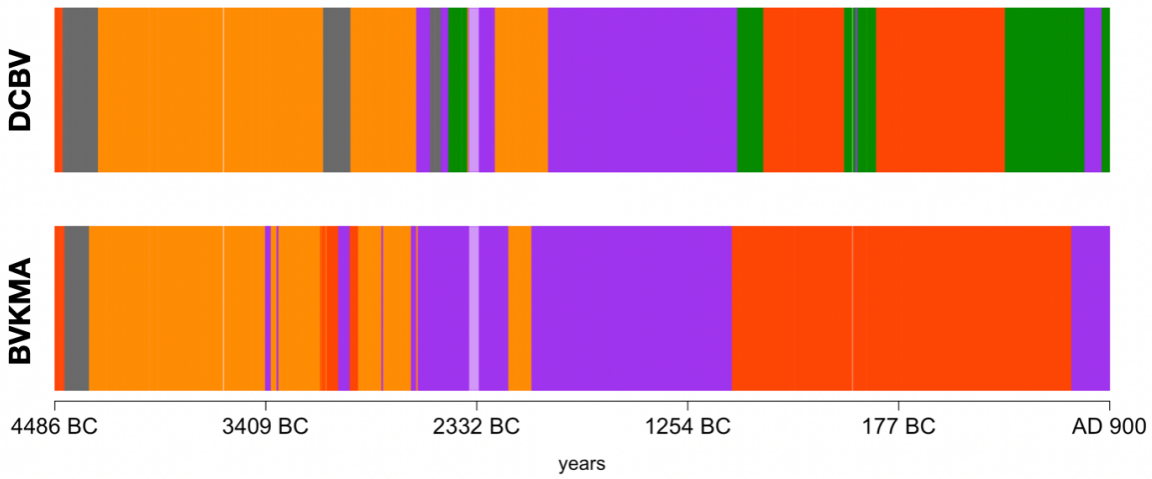


Figure A.1: Climate labels according to DCBV (top) and BVKMA (bottom) when $K = 5$ and $L = 100$

The computational cost is not the only reason why BVKMA has not been selected as the final model. Consider Figure A.2: each column represents the content of a specific climate cluster according to the three lakes and, thus, each row shows the interpretation of the four clusters according to Kassjön (first row from the top), Korttajärvi (middle row) and Nautajärvi (bottom row). Clearly, the four climate types don't seem to contain particularly different sedimentation patterns, considering the analogous shapes of the climate representatives. This is due to the fact that BVKMA follows the interpretation of climate as the *average trend* of weather over a long period of time, since each cluster (and cell) is represented by the average of the aligned functions inside it. However, each climate lasts for hundreds of years: indeed, the number of years assigned to the gray cluster by BVKMA is 124, the orange climate contains 1514, the purple one 1742 and finally 1891 consecutive years are assigned to the red-colored climate type. This entails that the functions shown in Figure A.2 are meant to represent the average behaviour

of thousands of years per cluster. As noted in Abramowicz et al. (2019), the simple pointwise mean is not guaranteed to be an efficient way of representing many consecutive years, since it tends to annihilate any small variations in the functions it represents. For example, let us imagine that, after the alignment, a portion of the signals contained in one climate (or cell) has an unimodal behaviour, while the remaining portion has mainly two peaks. The result of the pointwise average of these two tendencies within the same cluster (or cell) is not expected to properly attest to the unimodal behaviour of a portion of the signals. On the other hand, representing the cluster (or cell) with the frequency with which each weather type appears within the cluster (or cell) itself will allow to still take into consideration the unimodal tendencies, as long as the first-level clustering model is able to detect them.

The same analysis was conducted with $K \in \{3, 6\}$ and with $L \in \{50, 300\}$ to see the outcome of choosing high or low values of the two parameters. The conclusion of these experiments are similar to the ones presented in this section, regarding the limitation of the pointwise average to accurately depict diverse climate types.

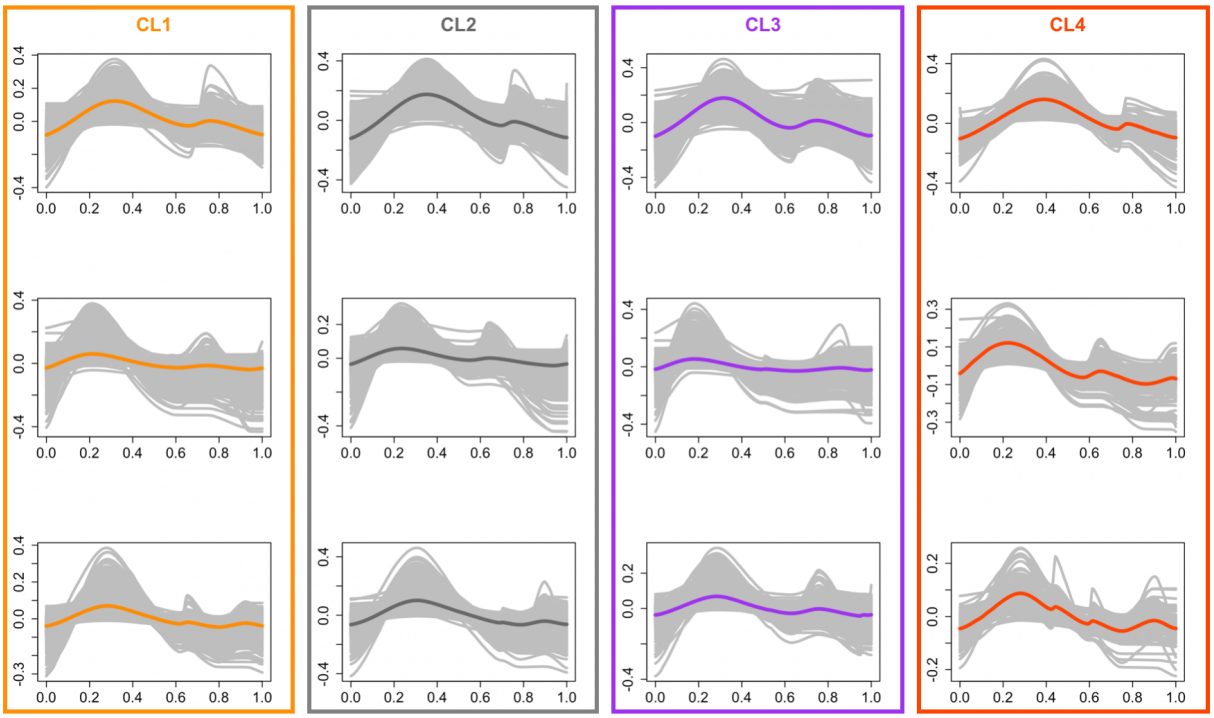


Figure A.2: Content of each climate cluster according to BVKMA when $K = 5$ and $L = 100$: the first row from the top in each cluster CLj pertains to Kassjön, the second to Korttajärvi and the third one to Nautajärvi.

A.2. Evidence of agreement in the weather types

As stated in the Introduction, previous works analyzed the climatological information contained in the signals of the three lakes, treating each of them independently of the other two. In particular, the results of Pozzari (2017) show that there is evidence of a common message in the lakes, based on an a posteriori analysis of the *local* climate types. The goal of this section is to show that it is possible to uncover this evidence using the first-level labels, namely analyzing the weather. This task is quite ambitious, considering the extremely local nature of weather and the fact that the weather types don't always match across the lakes (see Section 3.1). Hence, we devised a technique which does not rely on comparing the weather types of different lakes directly.
For each lake we followed this routine:

1. Divide the years into a suitable number M of equally-spaced and non-overlapping windows (see Figure A.3).
2. Represent the years within each window with a five dimensional vector, the j-th component of which is the proportion of times in which the j-th weather label of a specific lake appears within the cell. In this way, it is possible to obtain an $M \times 5$ dimensional dataset for each lake.
3. Compute the SVD decomposition of the (standardized) $M \times 5$ dimensional dataset for each lake, thus obtaining the first two principal components, their loadings and the scores.
4. Plot the first score for the M windows following their chronological order.

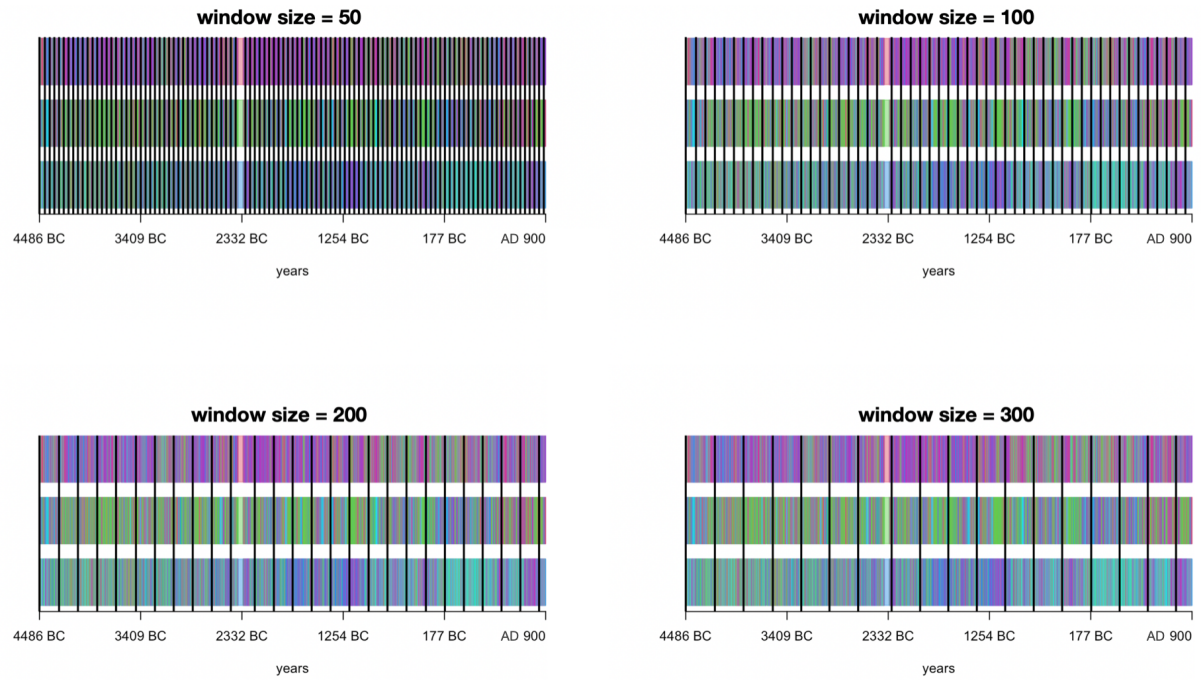


Figure A.3: Covering of the 5271 years with different choices for the window size. The role of the window size is similar to the temporal resolution parameter L.

This routine can be viewed as an application of PCA to a dataset constructed with a "moving window" approach, applied to the first-level weather labels for each lake. Thanks to the fact that the first principal component (PC1 for short) is a "mixture" of the frequency in which each lake-specific weather type appears in a window, it is possible to compare PC1 across the different lakes, even though their weather types don't necessarily match.

Figure A.4 shows that when the size of the windows is equal to 100 (to reproduce the proposed value of L for DCBV in Chapter 4) the first principal component explains approximately between 50% and 60% of the variability in the $M \times 5$ dimensional dataset for each lake. Moreover, Figure A.5 shows the value of the first score for the 52 windows when their size is set to 100 ($[5271/100] = 52$) as a function of the window indices[1]. The behaviours of the lake-specific first principal score don't appear to be "chaotic" with respect to the chronological order of the windows. Furthermore, the three curves show a similar trend across the lakes. The same can be stated conducting the PCA on the dataset constructed with a window size equal to 50 (see Figure A.6) and to 300 (see Figure A.7). Additionally, considering the first fifteen windows, the Finnish lakes showcase more similar trend with each other than with lake Kassjön. This is coherent with the results presented

[1]The indices follow the chronological order of the windows: for example, if window number j represents the years from 401 BC to 500 BC, then window number $j+1$ will span from 501 BC to 600 BC

in Section 5.1.4, in which we reached a similar conclusion inspecting the content of each climate type for lakes Korttajärvi and Nautajärvi.
The study conducted in this section has served the sole purpose of finding evidence of a shared message across the lakes at the "weather-level", using a proxy (namely, PC1) which does not rely on a specific weather type for each lake (which might not be present according to the other two), but rather on a mixture of them according to the loadings.

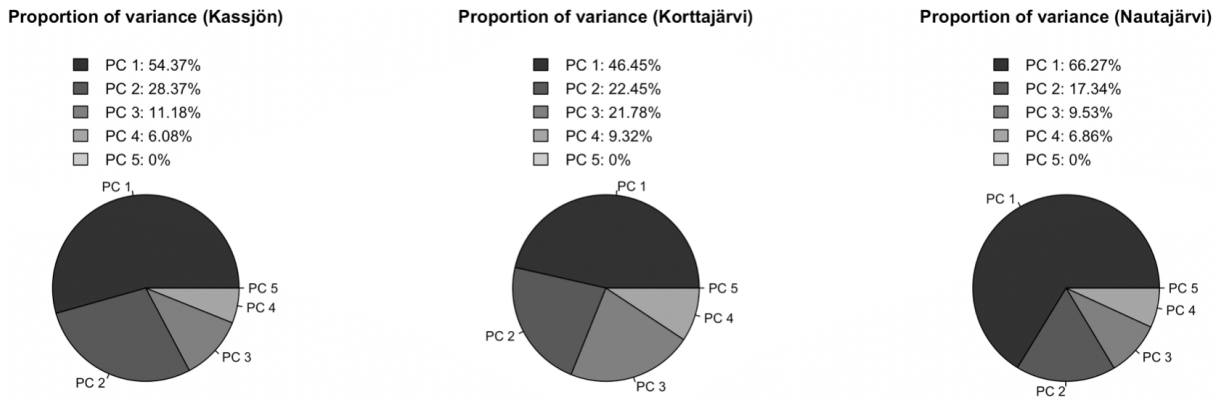


Figure A.4: Proportion of variance exaplained by the five PC when the size of the windows is equal to 100

Figure A.5: PC1 as a function of the window index, when the size of each window is 100

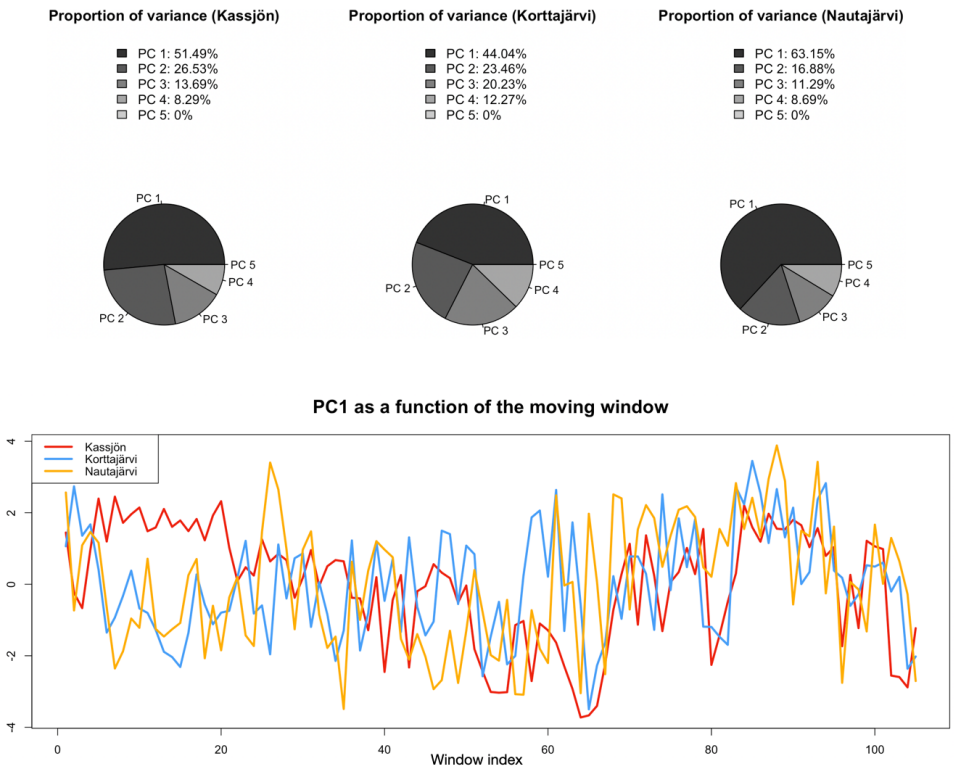


Figure A.6: Result of PCA, when the size of each window is 50

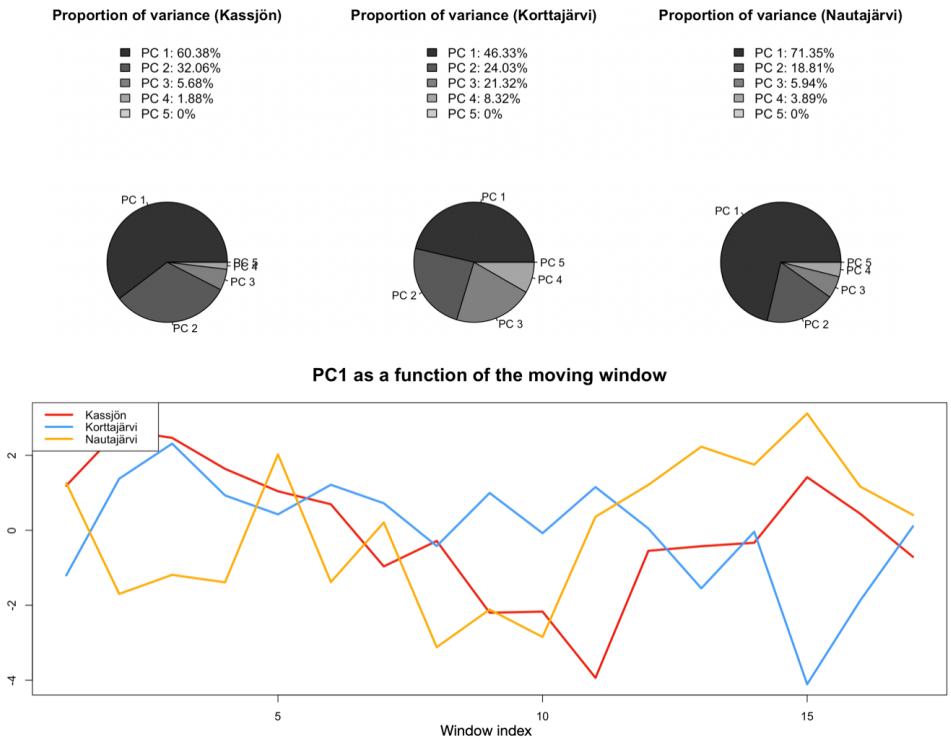


Figure A.7: Result of PCA, when the size of each window is 300

List of Figures

1.1 Example of phase and amplitude variability 6

2.1 Location of lake Kassjön, Korttajärvi and Nautajärvi 19
2.2 Satellite images of the three lakes 20
2.3 Varves' structure of lake Kassjön obtained with a digital signal processing video camera, *source:* Arnqvist et al. (2016) 22
2.4 Pre-processed Kassjön data 23
2.5 X-ray images of varved sediments 25
2.6 Pre-processed Korttajärvi data 25
2.7 Pre-processed Nautajärvi data 26
2.8 Plot of the time period spanned by each lake's data 27

3.1 Sample of five years across the three lakes to show the variability of lake Korttajärvi's signals 31
3.2 Boxplot of the cumulative absolute value of the first derivative $\left(\int_0^1 |\dot{f}(t)|dt\right)$ before (left) and after (right) the smoothing procedure applied to the signals from lake Kassjön. 31
3.3 Scatterplot of the peak location vs peak intensity (normalized in $[0, 1]$). The signals having a peak exactly in 0 and 1 could be wrongly delimited years and are thus removed before clustering. 32
3.4 Within- and between-cluster variability as functions of the over-alignment-penalizing parameter λ 34
3.5 Average similarity for each lake $\rho_\ell = \rho_\ell(k)$ as a function of the number of weather clusters $k \in \{3, 4, 5, 6\}$ 35
3.6 First-level clusters for lake Kassjönfor $k \in \{3, 4, 5, 6\}$ 36
3.7 Comparison between W1, W4 and W6 in lake Kassjön 36
3.8 First-level clusters for lake Korttajärvifor $k \in \{3, 4, 5, 6\}$ 38
3.9 First-level clusters for lake Nautajärvifor $k \in \{3, 4, 5, 6\}$ 38
3.10 Plot of the SRVF representatives for W2 and W4 regarding lake Nautajärvi when $k = 5$ (top) and $k = 6$ (bottom) 39

3.11 Final results of first-level clustering for each lake 39
3.12 First-level labels for each lake 41
3.13 Toy example of second-level clustering 48
3.14 Example of assignment of the first, second and third type 50
3.15 Example on how to compute the weights 52

4.1 Entropy plot for $K \in \{3, 4, 5, 6\}$ 57
4.2 Plot of yearly climate labels as a function of K 58
4.3 Supervised bagging Voronoi applied to each lake 60
4.4 Accuracy and recall for all L and $K \in \{3, 4, 5\}$ 61
4.5 One hundred simulations of SupBV applied to the second-level labels when $K = 5$ and $L = 100$ 62
4.6 Extending the timeline for Kassjön 64
4.7 Extending the timeline for Korttajärvi and Nautajärvi 64

5.1 Climate cluster representatives 66
5.2 Yearly climate labels ($K = 5$, $L = 100$) 66
5.3 Climate cluster representatives for Kassjön 67
5.4 Climate cluster representatives for Korttajärvi 68
5.5 Climate cluster representatives for Nautajärvi 70
5.6 Climate cluster comparison 72

A.1 Comparison BVKMA and DCBV 84
A.2 Content of each climate cluster according to BVKMA when $K = 5$ and $L = 100$ 85
A.3 Covering of the 5271 years with different choices for the window size 87
A.4 Proportion of variance exaplained by the five PC when the size of the windows is equal to 100 88
A.5 PC1 as a function of the window index, when the size of each window is 100 88
A.6 Result of PCA, when the size of each window is 50 89
A.7 Result of PCA, when the size of each window is 300 89

List of Tables

1.1 The first column presents the formalism of Vantini (2012) from the point of view of the space F, while the second one is from the point of view of the square-root velocity functions. 15

3.1 Estimated transition probabilities for the Markov Chains modeling the first level labels 41
3.2 Estimated sojourn times in each state (i.e. weather label) according to each lake 41
3.3 Permutational test to investigate time dynamics of weather 42

4.1 Accuracy and recall when $K = 5$ and $L = 100$ 62

List of Symbols

Variable	Description
(F, d)	metric space for the functions
W	space of warping functions
d_W	pseudo-metric on $F \times F$ induced by d
$\mathcal{F}$	quotient set of according to the relation $d_W(f_1, f_2) = 0$
$d_{\mathcal{F}}$	metric defined on $\mathcal{F} \times \mathcal{F}$
S	set of SRVF represenatives for the functions in F
d_{FR}	Fisher-Rao Riemannian metric
f	generic function in F
q	SRVF representative of a generic function in $f \in F$
n	number of years in the dataset ($n = 5271$)
$\{t_j\}_{j=1}^{200}$	equally spaced grid of points in which the functions are evaluated
k	number of first-level (weather) clusters
K	number of second-level (climate) clusters
$\{\boldsymbol{f}_i(t)\}_{i=1}^{n}$	collection of the three datasets
$f_i^1(t_j)$	i-th function in the dataset of lake Kassjön evaluated in t_j
$f_i^2(t_j)$	i-th function in the dataset of lake Korttajärvi evaluated in t_j
$f_i^3(t_j)$	i-th function in the dataset of lake Nautajärvi evaluated in t_j
φ_j	template of the j-th weather type in the first-level clustering
ψ_j	SRVF representative for the template of the j-th *weather* type
Wj	subset of the years assigned to the j-th *weather* cluster
CLj	subset of the years assigned to the j-th *climate* cluster
ℓ	index used to indicate the lakes
$\{\bar{\boldsymbol{g}}_c^b\}_{c=1}^{N}$	set of cell representatives in the b-th bootstrap iteration
$\{X_t^\ell\}$	Markov Chain modeling the weather labels pertaining to lake $\ell \in \{1, 2, 3\}$
$\{\boldsymbol{g}_i\}_{i=1}^{n}$	triplet of first-level labels (one for each lake)
g_i^ℓ	first-level (weather) label assigned to year i according to lake ℓ

Variable	Description
B	number of bootstrap iterations
S_0	one-dimensional lattice of years from 4486 BC to AD 900
N	number of nuclei in a Voronoi tessellation
L	temporal resolution parameter for a Voronoi tessellation

Acknowledgements

This thesis has been developed in collaboration with professor Konrad Abramowicz from the Department of Mathematics and Mathematical Statistics at Umeå University, in Sweden. I am extremely grateful to him for his support and encouragement throughout my six month stay at the department. I would also like to thank Alp Yurtsever and all the people I met at Umeå University, who helped me by sharing their ideas on how to improve this work.