

Clustering misaligned dependent curves applied to varved lake sediment for climate reconstruction

Konrad Abramowicz¹, Per Arnqvist¹, Piercesare Secchi², Sara Sjöstedt de Luna¹, Simone Vantini², Valeria Vitelli³

Abstract In this paper we introduce a novel functional clustering method, the Bagging Voronoi K-Medoid Alignment (BVKMA) algorithm, which simultaneously clusters and aligns spatially dependent curves. It is a nonparametric statistical method that does not rely on distributional or dependency structure assumptions. The method is motivated by and applied to varved (annually laminated) sediment data from lake Kassjön in northern Sweden, aiming to infer on past environmental and climate changes. The resulting clusters and their time dynamics show great potential for seasonal climate interpretation, in particular for winter climate changes.

Keywords Functional data · Clustering · Dependence · Misalignment · Sediment data

1 Introduction

Concern for global warming and future climate changes, and the impact of human activities related thereto, has increased the urge to understand how the climate has varied over the last centuries and millennia. For instance, climate models used to predict future climate (Pachauri et al. 2014) rely on a good understanding of past climate variability. Since instrumental meteorological records of the climate are scarce prior to mid nineteenth century it becomes necessary to rely on natural climate archives such as tree-rings, ice cores, corals, as well as lacustrine and marine sediments to reconstruct climate at long-term scales (see, e.g., Mann et al. 2008). Information about past changes in the seasonality of the climate and in particular winter climate is still scarce. This is a significant shortcoming, given how the recent climate change has affected different seasons unequally, (e.g., Beniston 2005), not the least at higher latitudes.

Varved (annually laminated) lake sediments have the potential to play an important role for understanding past seasonal climate with their inherent annual time resolution and within-year seasonal patterns. Several attempts to produce high-resolution (annual) reconstruction of climate from varve properties include work on Swedish and Finnish lakes by Petterson et al. (1999), Tiljander et al. (2003), Ojala and Alenius (2005), Ojala et al. (2008) and Petterson (1999). However, none of them have fully taken into account the information contained in the shape of the seasonal patterns. Such data can be seen as a collection of functions (seasonal sediment profiles) whose changes through time it is of interest to study with respect to climate. Functional data analysis (see, e.g., Ramsay and Silverman 2005) provides successful statistical tools for environmental, weather, and climate studies. For instance,

Published online: 14 July 2016

✉ Simone Vantini
simone.vantini@polimi.it

Konrad Abramowicz - konrad.abramowicz@umu.se

Per Arnqvist - per.arnqvist@umu.se

Piercesare Secchi - piercesare.secchi@polimi.it

Sara Sjöstedt de Luna - sara.sjostedt.de.luna@umu.se

Valeria Vitelli - valeria.vitelli@medisin.uio.no

¹ Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden

² MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy

³ Department of Biostatistics, Oslo Center for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway

with respect to some recent applications, functional kriging techniques have been used to analyze particle-size curves pertaining to ground samples collected in an aquifer (Menafoglio et al. 2014), the estimated distribution of maximal temperature (in the month of January) recorded in 29 Canadian weather stations (Salazar et al. 2015), and daily PM10 concentrations recorded by a monitoring network in Northern Italy between October 2005 and March 2006 (Ignaccolo et al. 2014). Functional-on-functional regression techniques have been used to model depth measures of ocean temperature as functions of salinity and fluorescence depth measures recorded in 1585 stations in Southern Atlantic (Fernández-Pascual et al. 2015). Functional random fields have been used to analyze oxygen concentrations in sedimentary columns (Dabo-Niang et al. 2010) and basal area growth curves for forest trees (Comas et al. 2013). Finally, functional clustering techniques have been used to analyze lake surface water temperature profiles along 5 years for 256 American lakes (Finazzi et al. 2015) and the yearly seasonal patterns of varved lake sediment of lake Kassjön in northern Sweden (Arnqvist et al. 2016).

In particular, in Arnqvist et al. (2016) the aim was to cluster the seasonal patterns into homogeneous groups, that correspond to different yearly weather types, characterizing the area of lake Kassjön. Their results revealed a clustering structure and it was also noted that the data are affected by horizontal misalignment, mainly imputable to the different sedimentation rates within and across different years which can act as confounding factors when applying clustering techniques, see, e.g., Gaffney and Smyth (2004).

In climatology, focus is typically on the long-term weather trends, i.e., climate, rather than weather fluctuations. In order to capture climate evolution in data such as the Kassjön sediment, there is a need for a functional clustering method that is able to jointly deal with the temporal dependence, the misalignment, and the presence of clusters that characterize the underlying seasonal patterns.

There is a large corpus of consolidated literature dealing with clustering, alignment, and spatial dependence of functional data which is cleverly summarized in top journal special issues and review articles. Nevertheless, within this literature these three sources of variability have mostly been tackled separately, with very few exceptions jointly dealing with more than one. A common feature of these latter works is to point out the importance of taking into account all sources of variability which functional data are likely to be affected by, to achieve an accurate and realistic modeling of the phenomenon under investigation.

To be more specific, methods for simultaneous clustering and alignment of functional data (without addressing potential spatial dependence between functions) have been proposed by, e.g., Liu and Müller (2004), Gaffney and

Smyth (2004), Sangalli et al. (2010a, b, 2014). Methods for clustering spatially dependent functional data (without addressing potential misalignment of the functions) have been proposed by, e.g., Ignaccolo et al. (2008), Romano et al. (2010, 2015), Secchi et al. (2011, 2013), Giraldo et al. (2012) and Menafoglio et al. (2016). We are not aware of any methodological work for the alignment of spatially dependent functional data, nor of any work for simultaneous clustering and alignment of spatially dependent functional data.

In this paper, we propose a novel method, the Bagging Voronoi K-Medoid Alignment algorithm (BVKMA), that jointly handles clustering, misalignment, and spatial dependence of functional data. Up to our knowledge, this method is the first proposal in the literature that jointly deals with these three sources of variability. Furthermore, with the purpose of capturing the underlying different climate regimes, we use this new method to re-analyse the seasonal patterns of the sediment data from Kassjön firstly analyzed in Arnqvist et al. (2016) by means of a functional K-mean clustering (applied to landmark registered curves, without taking into account spatial dependence). The proposed method is purely metrical thus exploring the three sources of variability in a nonparametric fashion, avoiding both parametric distributional assumptions pertaining to amplitude and phase variability and a parametric modeling of the dependency structure. Moreover it allows many different families of warping functions to address the misalignment, and it is flexible with respect to the choice of distance measures used to quantify the similarity of curves and to the number of clusters.

In Sect. 2 we describe the Lake Kassjön sediment data and the undertaken preprocessing. Section 3 describes the functional clustering methods used, with emphasis on the new method that simultaneously takes into account dependence and misalignment. In Sect. 4 we present the result of applying the methods to the Kassjön sediment data. A simulation study is presented in Sect. 5 and a discussion of the results and methods is found in Sect. 6.

2 Data and preprocessing

The varved sediment to be analyzed is taken from the bottom of Lake Kassjön (63° 55' N, 20° 01' E), situated in northern Sweden, Petterson (1999). A subsequence of the sediment is shown in Fig. 1. These patterns reflect to a large extent weather conditions and internal biological processes in the lake the year the varve was deposited at the bottom of the lake. The annual varve patterns have the following origin. During spring, in connection to snow melting and spring runoff, minerogenic material is transported from the surrounding landscape into the lake via

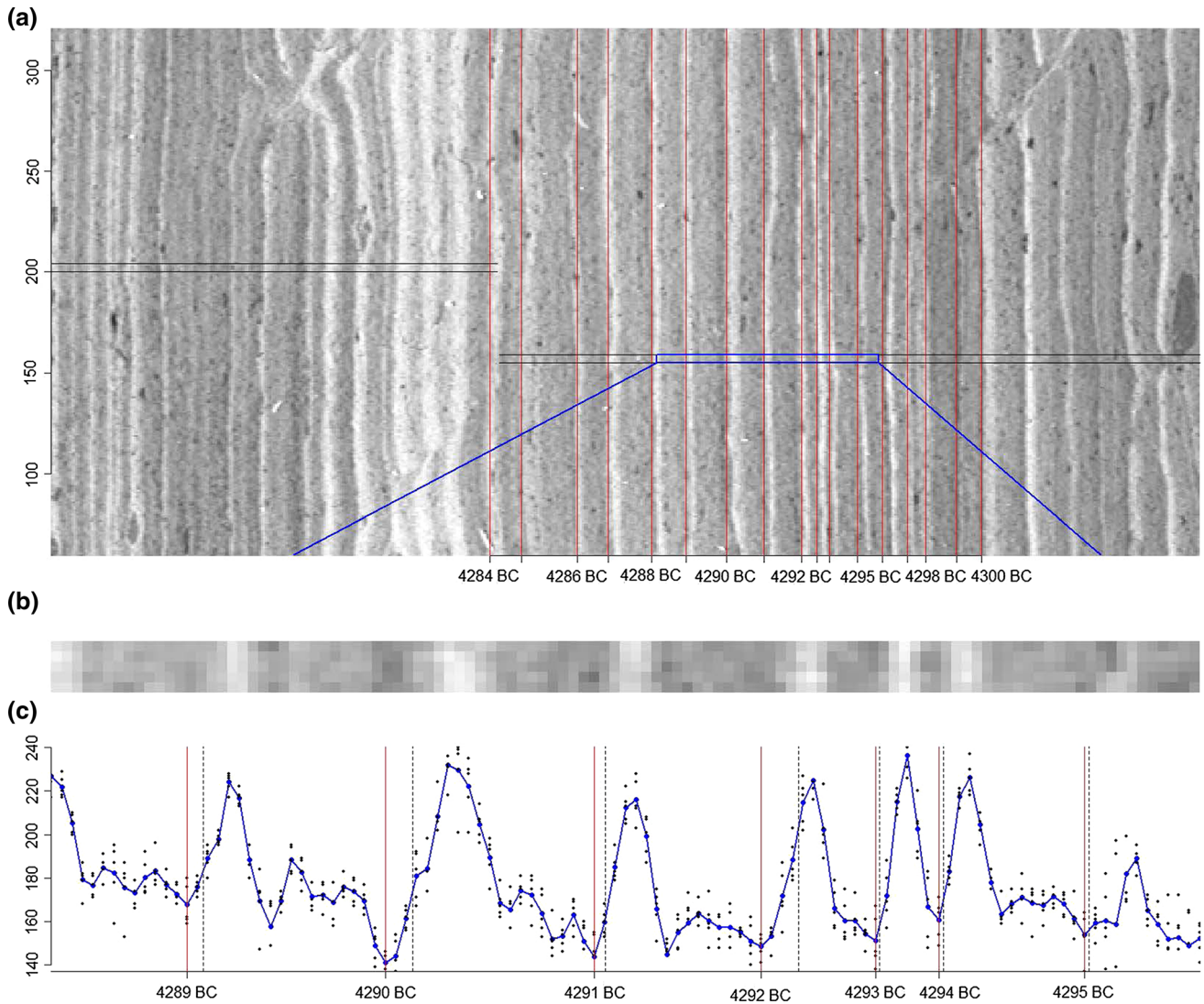


Fig. 1 **a** Annual sediment from lake Kassjön. **b** Data to be analyzed is based on slices of five pixels width selected from representative parts of the sediment. **c** Grey-scale values for the slice in **b** together with the mean grey-scale values (solid line) of the 5 pixels for each

time point. The manually determined yearly delimiters (black dotted lines) have been horizontally shifted 1–4 steps to the darkest neighboring value (solid red lines)

four small streams, which gives rise to a bright colored layer (high greyscale values). During summer, organic material produced mainly by the lake organisms, sinks to the bottom and creates a darker layer (lower greyscale values). Finally, under the ice cover during winter, fine organic material is deposited, resulting in a thin blackish layer (lowest greyscale values). The minerogenic input reflects the intensity of the spring run-off, which is dependent on the amount of snow accumulated during the winter, and hence the minerogenic input reflects the long-term record of variability in past winter climate, Petterson et al. (2010). The seasonal profiles have the potential to carry important weather information on a seasonal level: A pronounced spring peak (in terms of gray scale values) may indicate a winter with rich amounts of snow, a low spring

peak a winter with less snow. A substantial flatter part after the spring peak would correspond to a thick organic sediment layer perhaps indicating a warmer summer. Peaks appearing after the first (spring) peak may indicate fall storms with heavy rain.

The annual seasonal patterns of the sediment, covering approximately 6400 years, were recorded as greyscale images, following the method described by Petterson et al. (1993, 1999). The greyscale values range from 0 to 255, where 0 and 255 corresponds to black and white, respectively. The raw data set is a series of averages of five-pixel slices selected from representative parts of the varved sediment images, cf. Figure 1b-c, Petterson et al. (1993, 1999). Varve delimiters were initially set manually by two experts, studying the sediment core using stereo

microscopes. A varve is defined from the beginning of one spring layer to the beginning of the next, since the shift from the winter layer to the spring layer is the sharpest transition in varves of this type (Pettersen et al. 1993). The varve delimiters should thus correspond to the thin blackish layer produced when ice covers the lake. When converting the manually determined delimiters to the greyscale values, some were horizontally shifted (1–4 pixels), to make sure they corresponded to the darkest pixel-values in the neighborhood. In this way, the final raw data was obtained, being composed by a time series of greyscale vectors (of different lengths) associated to years from 4486 B.C. to A.D. 1901. Of the 6388 varves, 62 of them had no greyscale values recorded and were therefore treated as missing, see Arnqvist et al. (2016) for details.

The yearly data consist of several components which may be important carriers of (weather) information, e.g., varve thickness, mean greyscale value, and the corresponding within-year dynamics around it. The yearly dynamics gives insight into the seasonal variation and hence allows to draw inference on a more detailed level, removing the effect of the potentially confounding mean greyscale level, see Arnqvist et al. (2016). Therefore the raw sediment data is normalized by centering each year around zero. Without loss of information the yearly time scale is converted to a reference one by uniformly distributing the time instances on the interval $[0, 1]$ (such that for each year the first time instance is associated to 0 and the last one to 1), which corresponds to a first affine time transformation.

Since the greyscale value of each pixel provides a local evaluation with error of a continuous underlying (sedimentation) process, a functional description of these data comes natural. Starting from the normalized final raw sediment data, we build the functional data set by smoothing (separately for each year) the yearly normalized data through a penalized B-spline approach using a least squares fitting criterion and a penalization based on the squared L^2 norm of the second derivative (see, Chap. 5, Ramsay and Silverman 2005). We use cubic splines with 32 basis functions and equally spaced knots on the interval $[0, 1]$. The penalty constant was set to $\lambda = 0.000140625$, for more details see Arnqvist et al. (2016).

3 Clustering methods for misaligned and/ or dependent functional data

We propose a novel functional clustering method called the Bagging Voronoi K-Medoid Alignment (BVKMA) algorithm, which jointly handles misalignment and (temporal) dependence in the functional data. This method integrates

two functional clustering methods, the K-Medoid Alignment method (KMA) and the Bagging Voronoi K-Medoid strategy (BVKM). We first describe the main ideas of these two methods followed by a presentation of the BVKMA method.

3.1 K-Medoid Alignment: clustering misaligned functional data

The functional K-Medoid Alignment method introduced by Sangalli et al. (2010a, b) is a generalization of the functional K-Medoid clustering algorithm (see, e.g., Tarpey and Kinader 2003), which jointly aligns and clusters a set of observed functions. This method is in fact obtained as the integration of a continuous alignment method based on the Procrustes procedure into functional K-Medoid. The permitted warping functions in the algorithm belong to a warping group \mathcal{W} , which for example can be a group of positive slope affine transformations.

The KMA is an iterative method which at each iteration performs the following steps:

- (i) the *medoid identification step*, in which cluster medoids are chosen as the curves in each cluster which are closest to all the other aligned curves in the same cluster,
- (ii) the *cluster assignment and alignment step*, in which each curve in the sample is assigned to the cluster whose medoid is the closest, after being aligned to each medoid using an optimal warping function in \mathcal{W} ,
- (iii) and finally the *normalization step*, which is performed to ensure that the average warping undergone by curves assigned to each cluster is the identity transformation.

The results of this procedure are: a cluster assignment, an estimated warping function (describing the misalignment) for each curve in the sample, and a set of K estimated medoids.

3.2 Bagging Voronoi K-Medoid: for dependent functional data

An extension of the functional K-Medoid clustering algorithm that can deal non-parametrically with dependent functions is the Bagging Voronoi K-Medoid algorithm. This procedure has originally been proposed in Secchi et al. (2011, 2013) for dealing with spatially dependent functional data, indexed by the sites of a spatial lattice.

In particular, this method is based on aggregating the results obtained from B random bootstrap replicates of the same analysis. This is the so called *Bootstrap Phase* of the

method, and each bootstrap replicate is composed of the following three steps:

- (i) *generation of a random Voronoi tessellation* over the considered lattice. This means sampling a random set of sites (years) to be the nuclei of the tessellation, and then assigning each of the other sites to the closest nucleus. For a 1-dimensional lattice of years, a Voronoi tessellation is a random set of intervals of time over the years.
- (ii) *identification of a functional representative* for each element of the tessellation. The functional local representative (i.e., the medoid of the data associated to the same tessellation element) summarizes the information carried by all functional data indexed by sites (years) belonging to the same element of the tessellation.
- (iii) *clustering of the local representatives* Once the sample of functional local representatives is obtained, a K-Medoid functional clustering procedure is applied to obtain a final classification. For each Voronoi map all sites belonging to the same tessellation element get the same cluster label as its local representative.

These three steps are repeated B times. Thus, for each year, a frequency distribution of cluster assignments along the B replicates is provided. This is part of the so called *Aggregation Phase* of the method. The computation of the frequency distribution of cluster assignment along the bootstrap replicates is made after a relabelling procedure is applied to match labels along replicates. For each year, the final label is the result of a majority vote on cluster assignments along the replicates. The functional representatives of the final clusters are then constructed as their corresponding functional medoids. The whole procedure is visually sketched for a one dimensional lattice in Fig. 2.

Moreover, an entropy criterion (originally proposed in Secchi et al. 2013) can be used to inspect the variability of the frequency distribution of cluster assignments, thus quantifying the uncertainty associated to the classification. This measure serves as a practical tool in selecting the appropriate number of nuclei for the Voronoi tessellations for the application at hand.

3.3 Bagging Voronoi K-Medoid Alignment: for misaligned dependent functional data

The novel method we here propose jointly handles the misalignment and (temporal) dependence in functional data, and it is obtained by merging the K-Medoid Alignment and the Bagging Voronoi K-Medoid strategies, which separately tackle the two issues of interest, respectively. Moreover, it provides important benefits with respect to the

BVKM, since misalignment can act as a confounding factor when accounting for dependence alone.

The building blocks of the method are based on BVKM, but both the bootstrap and the aggregation phase are modified to deal with misalignment along the lines depicted by KMA. The basic modifications in the bootstrap phase are

- each functional representative is obtained by applying 1-medoid alignment to the functional data belonging to the corresponding tessellation element;
- the functional representatives are clustered via KMA.

The modifications to the aggregation phase are less demanding, and will be explained in details in the following. Without loss of generality we describe the algorithm for (dependent) functions along a one-dimensional grid indexed by $j = 1, \dots, N$. Let $f_j(t), t \in \mathcal{D}_j, j = 1, \dots, N$, denote the observed dependent functions.

For the Bagging Voronoi procedure, set the number B of bootstrap replicates, the number n of elements in the Voronoi tessellation, the metric d used to compute distances among functional data and the number K of clusters considered in the clustering procedure.

Then, the *Bootstrap phase* of the BVKMA algorithm is articulated as follows. For $b = 1, \dots, B$, replicate the steps:

- (i) A set of nuclei $\{\mathbf{Z}_1^b, \dots, \mathbf{Z}_n^b\}$ is randomly generated among the indices $1, \dots, N$ according to a uniform discrete distribution on the first N integers, i.e. for $i = 1, \dots, n$, $\mathbf{Z}_i^b \sim \mathcal{U}(1, \dots, N)$ without replacement. Then, the b -th random 1-dimensional Voronoi tessellation (i.e. a set of random intervals) $\{V_i^b\}_{i=1}^n$, is obtained by assigning each index (site) $j = 1, \dots, N$ to the nearest nucleus \mathbf{Z}_i^b .
- (ii) Given the tessellation, for $i = 1, \dots, n$ the local representative g_i^b , corresponding to the i -th element of the tessellation V_i^b , is computed as the single medoid resulting from 1-medoid Alignment applied to the dataset $\{f_{j_i}(t)\}_{j_i \in V_i^b}$.
- (iii) Finally, the set of functional representatives $\{g_1^b, \dots, g_n^b\}$ is clustered via KMA. The label obtained for each functional representative g_i^b , for $i = 1, \dots, n$, is then assigned to all sites belonging to the corresponding tessellation element.

At each replicate in the Bootstrap phase, we save the cluster labels for each of the N sites. The classifications obtained along replicates in the Bootstrap phase are then relabelled, analogously to the BVKM algorithm.

We now turn to the *Aggregation phase*. Let C_j^b be the label of site j in the b -th bootstrap replicate, for $j = 1, \dots, N$ and $b = 1, \dots, B$. Then, similarly to the BVKM method, the frequency distribution of cluster assignments

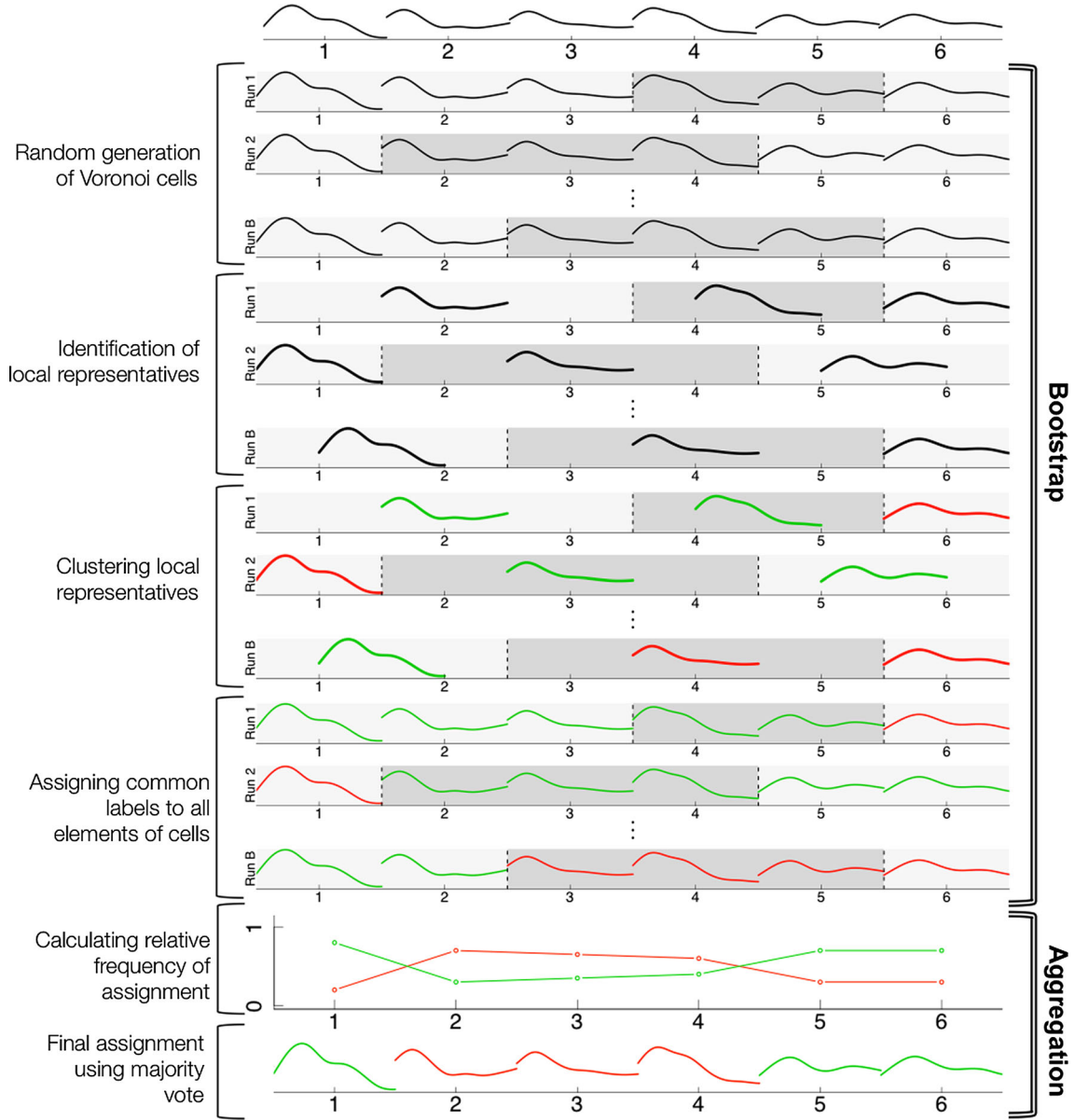


Fig. 2 A graphical scheme of the Bagging Voronoi K-Medoid strategy on a one dimensional lattice, with $K = 2$

along the B replicates is computed for $j = 1, \dots, N$ as $\pi_j^k = \#\{b \in \{1, \dots, B\} : C_j^b = k\} / B, \forall k = 1, \dots, K$. A final classification c_1, \dots, c_N is obtained by a majority vote with respect to the K clusters, i.e., $c_j = \operatorname{argmax}_{k=1, \dots, K} \pi_j^k$ for $j = 1, \dots, N$. The functional representatives of the final clusters are then constructed via 1-medoid Alignment within each of the final clusters.

Note that, for a given number of nuclei, n , the expected number of sites within each Voronoi element is $L = N/n$. Hence, we can interchangeably use L or n as a parameter in our method (for a given sample size N). The BVKMA analysis is robust with respect to moderate variations of the average length L of the elements of the random Voronoi

maps. However, an extreme reduction of L leads—as expected—to results similar to the ones provided by KMA while an extreme enlargement leads to non interpretable clusters (see Sect. 3.2 in Secchi et al. 2013). When the curves are (spatially) dependent, we typically gain information by taking into account information from neighboring curves, thus expecting L to be larger than one. If no dependence exist between curves, L is expected to be close to one. Tuning L is hence an important issue when running the BVKMA. If the problem at hand does not come with a natural value for L , we can choose the optimal value for L as the one providing a more distinct assignment of years to clusters, as suggested in Secchi et al. (2013). In

more detail, for each given L we can compute the site-by-site entropy, $E_j = -\sum_{k=1}^K \pi_j^k \ln \pi_j^k, j = 1, \dots, N$, of the bootstrap distribution of cluster labels which the site has been assigned to, and average it across the grid. The L corresponding to the smallest average entropy, $\bar{E} = \sum_{j=1}^N E_j/N$, indicates the best choice of L .

As described above, the Bagging Voronoi K-Medoid Alignment algorithm deals with clustering, misalignment, and dependence in a flexible non parametric fashion, avoiding both parametric distributional assumptions pertaining to amplitude and phase variability and a parametric modeling of the dependency structure. With respect to the clustering, the search for clusters is indeed driven by the selected functional metric (or similarity) which can flexibly focus on different data features of particular interest for the application at hand. The alignment process is also performed on a purely metric basis derived from the functional metric above and from a family of warping functions which is only required to have a group structure with respect to composition and to satisfy an invariance property with respect to the functional metric (i.e., Vantini 2012). Finally, the way in which spatial/temporal dependence is exploited to empower the analysis is exclusively related to the seed sampling strategy and to the spatial/temporal metric used to measure spatial/temporal distances between data observation points. The sampling strategy and the spatial/temporal metric can both be adapted to locally manage the bias-variance tradeoff in the computation of the local representatives (i.e., Secchi et al. 2011) and to locally model the local range and anisotropy of the dependence structure with respect to some application-related knowledge.

4 Application to the analysis of varved sediment data

In this section we analyze the sediment data from lake Kassjön. We apply the BVKMA algorithm (described in Sect. 3.3) with the aim of capturing climate variation, and the KMA algorithm (described in Sect. 3.1) to describe (yearly) weather regimes based on the seasonal patterns. For the sediment data, a one dimensional grid is naturally formed by the corresponding years. The functional data consists of $N = 6388$ curves $f_j(t), t \in [0, 1], j = 1, \dots, N$ describing the within-year greyscale dynamics around the annual greyscale mean.

The amplitudes of the peaks bring important information about the weather and therefore a choice of an L^2 -based measure comes natural to quantify the similarity between functions. We choose to use positive slope affine transformations as the group \mathcal{W} of warping functions. We expect the

group to be flexible enough to deal with the most important components of the misalignment in our data. More specifically, to measure the distance between two functions, $f_1(t), t \in \mathcal{D}_1$ and $f_2(t), t \in \mathcal{D}_2$, with $\mathcal{D}_1 \cap \mathcal{D}_2 \neq \emptyset$, we use the following normalized L^2 - distance

$$d(f_1(\cdot), f_2(\cdot)) = \frac{1}{\sqrt{\lambda(\mathcal{D}_1 \cap \mathcal{D}_2)}} \left(\int_{\mathcal{D}_1 \cap \mathcal{D}_2} (f_1(t) - f_2(t))^2 dt \right)^{1/2}, \quad (1)$$

where $\lambda(\cdot)$ is the Lebesgue measure. The use of the normalizing constant in (1) is justified by the following invariance property of this measure with respect to affine warping functions (e.g., Vantini 2012): for $f_1(t), t \in \mathcal{D}_1$ and $f_2(t), t \in \mathcal{D}_2$ and $h(t) = a + bt, a \in \mathbb{R}, b \in \mathbb{R}^+$, we have $d(f_1(\cdot), f_2(\cdot)) = d(f_1(h(\cdot)), f_2(h(\cdot)))$.

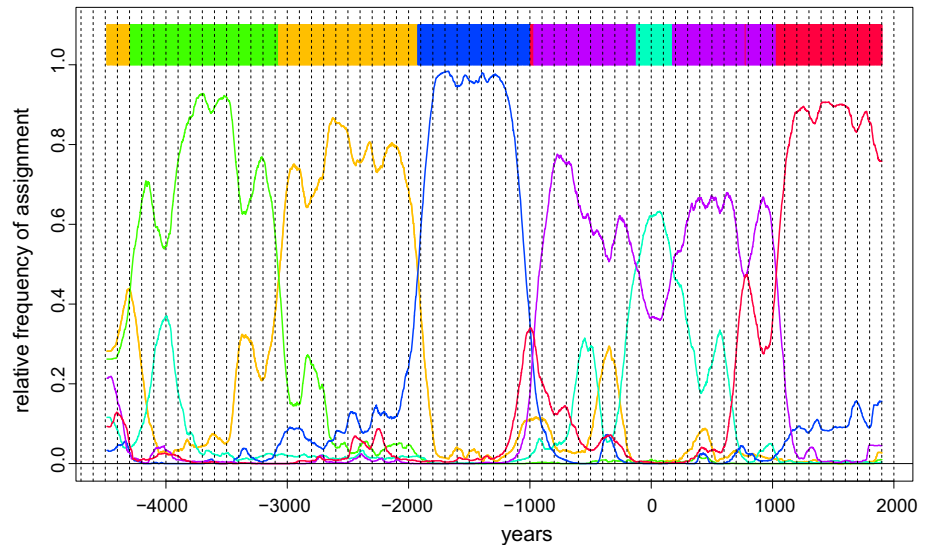
4.1 The BVKMA and KMA clusters

To identify different climate-types and their corresponding climatic periods with the BVKMA algorithm, we choose the number of clusters to be $K = 6$. The expected length (number of years) of the elements of the Voronoi maps is fixed to $L = 200$ and the number of bootstrap replicates is chosen to be $B = 1000$. A discussion about these parameter choices is found in Sect. 4.2. In the aggregation phase, the matching procedure used is based on Stephens's relabelling algorithm (see Stephens 2000). All used algorithms have been implemented in the R programming language (R Core Team 2015).

Using the L^2 - distance (1) and affine positive slope warping functions, the BVKMA algorithm was run on the Kassjön data. The aggregation phase relative frequencies are presented in Fig. 3. For most of the years, there was a clear dominating cluster label yielding the majority vote. Exceptions can be seen for years close to transitions between (majority vote) clusters, and for the initial few hundred years. During that initial period the lake was formed as a marine isolation basin by isostatic land uplift, and the sediment dynamics therefore mainly reflects the creation process of the lake.

In Fig. 4, the obtained BVKMA cluster medoids (C1–C6) are presented. Since the medoids are constructed via 1-medoid alignment of the annual profiles within each of the final clusters, they are easily interpreted as the typical (annual) weather curves in the corresponding BVKMA clusters. The obtained clusters are associated with six time periods. First a period of 1200 years (4300 BC, 3100 BC) characterized by medoid C1 with a sharp high spring peak. It is followed by 1150 years characterized by medoid C2 with a less pronounced spring peak and a flatter part thereafter, while the third period (1950 BC, 1000 BC), is represented by a low spring peak and a second lower peak

Fig. 3 The yearly relative frequencies of assignment over time obtained in the aggregation phase in BVKMA with $L = 200$. Different colors correspond to different clusters. The colored band represents the majority vote cluster assignment



in the flatter part thereafter (medoid C3). The time interval (1000 BC, AD 1000) excluding (150 BC, AD 150) is characterized by medoid C4 with a less pronounced spring peak. The shorter period of 300 years (150 BC, AD 150) is instead represented by a high spring peak (medoid C5). Finally, the period (AD 1000, AD 1900) is represented by a low and wide spring peak (medoid C6).

To gain additional insight into the structure and interpretation of the BVKMA clusters, we further investigate the frequencies of distinct yearly weather types, as defined by the KMA algorithm, within the corresponding BVKMA (climate) clusters. We thus group the seasonal patterns using the KMA methodology, with $K = 6$ clusters, the L^2 -distance (1) to quantify the distance between functions, and the affinities with positive slope as the group of warping functions. See Sect. 4.2 for a discussion of the choice of the number of clusters. The resulting six cluster medoids, presented in plots W1–W6 in Fig. 4, are characterized by high, medium or low spring peak, with or without a second minor autumn peak. We distinguish between KMA clusters with high sediment content-variation (such as clusters W2, W3, W6) and clusters which reflect stable, unchanged content of the sediment (cluster W4). The temporal distribution of the six KMA clusters appears noisy as opposed to the BVKMA clusters, see Fig. 5. It is consistent with the fact that annual weather varies substantially between neighboring years, whereas climate changes more slowly. A closer look at the temporal distribution of the KMA clusters reveals that their occurrences are not uniformly distributed over time. This is what climatologists call the evolution of climate across centuries. Each climatic period would thus be characterized by a particular frequency distribution over the different weather-types which is instead unusual in other climatic periods.

The relative frequency distribution of the KMA cluster labels for those years that occur in the corresponding BVKMA-cluster are illustrated in Fig. 4. We exemplify this additional information about the structure of the BVKMA-defined climatic periods by describing two contrasting groups. The first group, consisting of climate clusters C3 and C6, is characterized by a high frequency of the flat yearly profiles (KMA cluster W4) and low frequency of high spring peak years (KMA clusters W2, W3 and W6). During these climatic time periods, the largest proportion of the years thus corresponds to a low variation seasonal pattern, exhibiting lack of intense spring flood, indicating (milder) winters with small amounts of snow. On the other hand if we consider climate cluster C5, we see the opposite behaviour. The time period corresponding to C5 mainly consists of the high spring peak profiles and the flat stable profile is almost non-existent, thus indicating winters with large amounts of snow.

It is of interest to note that, the record of lake Kassjön show overall comparable patterns with the varved sediment from Nautajärvi (Finland) (Tiljander et al. 2003; Ojala and Alenius 2005; Arnqvist et al. 2016), indicating that the varved sediments are able to record a large-scale regional climate signal, and not only catchment processes of the Kassjön lake.

To confirm the climatic interpretations of the clusters we would like to link the varve record directly to regional meteorological observations, that started to be recorded in the late eighteenth century. This is however not straightforward since an increased amount of agricultural activities around the Kassjön lake has taken place in the last centuries, as well as ditching during the twentieth century (Segerström et al. 1984). This has affected the sediment transport and deposition processes considerably, possibly masking the climatic signal. We further make an attempt to

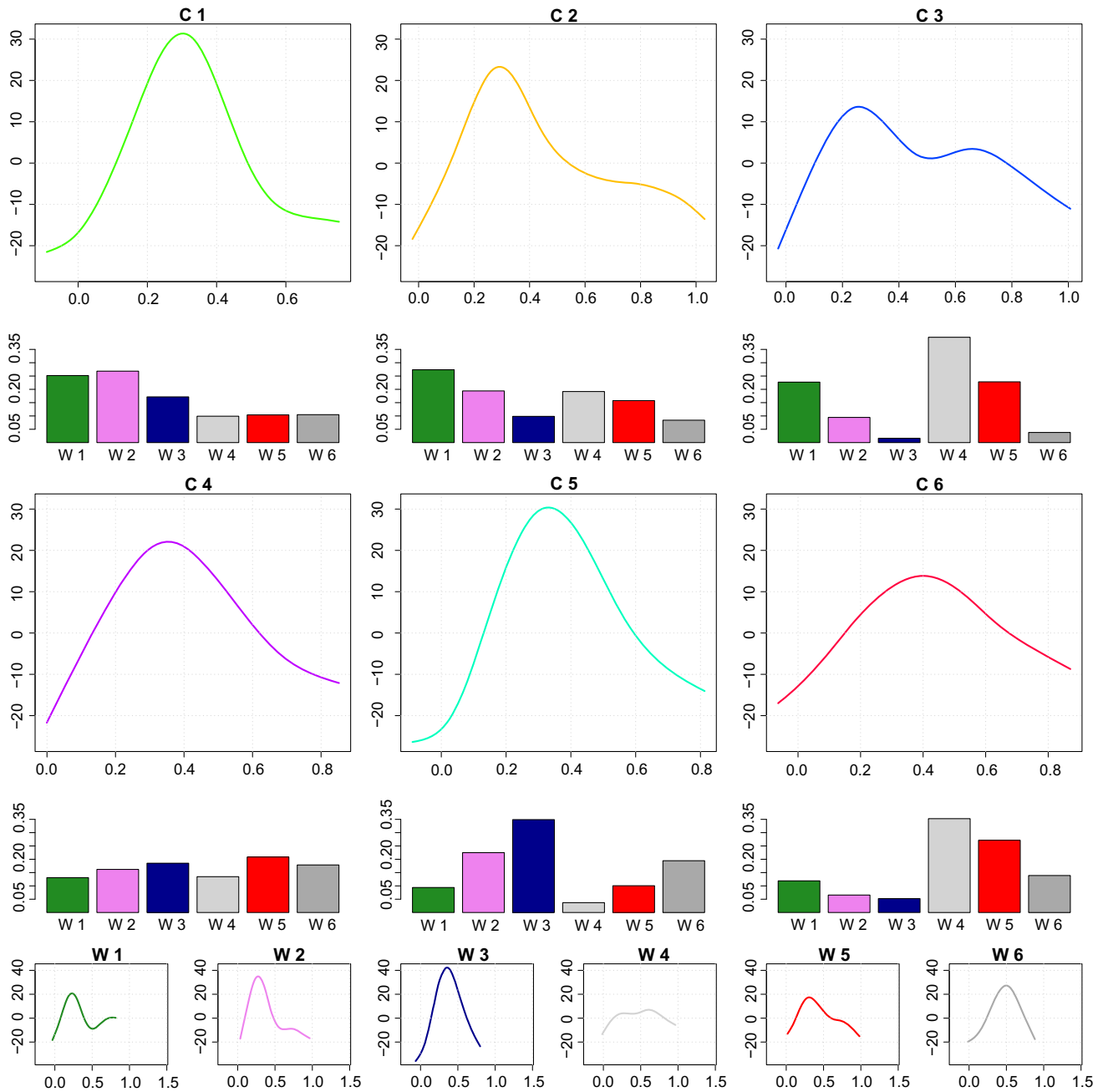


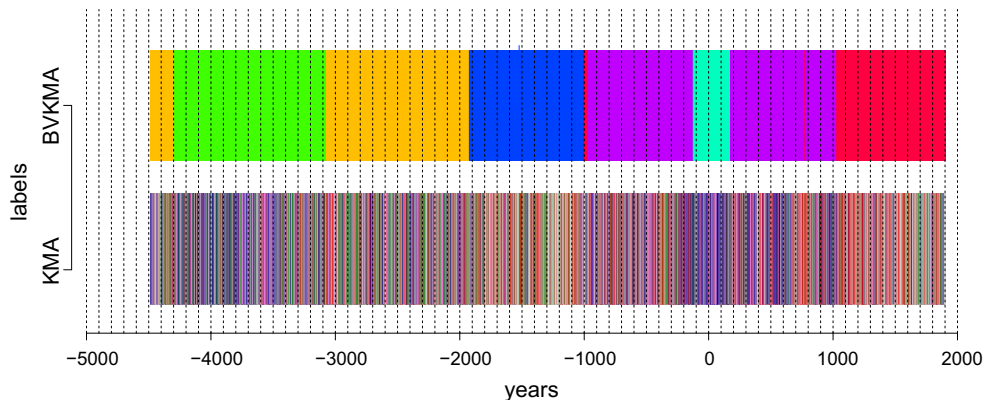
Fig. 4 The BVKMA cluster centroids (C1–C6) when using $L = 200$ together with the KMA centroids (W1–W6). The barplots present the relative frequencies of different KMA labels within each BVKMA

cluster. The colors of the BVKMA cluster centroids match the colors used in Fig. 3

relate the KMA clusters to an annual average winter temperature (January–April) reconstruction based on documentary sources of port activities in the Stockholm region, 1502–1892 by Leijonhufvud et al. (2010). There is a strong linear relationship between yearly average winter temperatures (January–April averages) in Stockholm and in Umeå, situated approximately 20 km from Kassjön. The correlation is 0.86, calculated on the recorded winter temperatures 1860–2009 provided by SMHI, with Umeå

winter temperatures being approximately 5 degrees lower than in Stockholm. Boxplots of the (reconstructed) Stockholm winter temperatures (1502–1892) within each KMA cluster is presented in Fig. 6. It is interesting to note that cluster W4 with flatter seasonal profiles tends to have warmer winter temperatures than KMA clusters with pronounced spring peaks. Still, there is substantial variability within clusters, parts of it likely attributable to potential uncertainties due to dating, reconstruction of Stockholm

Fig. 5 The time dynamics of BVKMA labels and KMA labels. The colors of the labels match the colors used in Fig. 4



temperatures and linking the Stockholm temperatures to temperatures in the Umeå region.

4.2 Parameter selection

In order to capture the full spectrum of seasonal dynamics of weather we need at least six clusters for the KMA algorithm: When choosing $K = 3$ clusters we obtained groups characterized by medoids with a large spring peak, a double peak and a flat profile. However, we then lack separation of the different spring peak amplitudes. A choice of four to five clusters still mixes the different amplitudes and the bi-modality, but with six clusters they clearly separate.

To decide the optimal number of clusters (climate regimes) for the BVKMA algorithm we used a stepwise procedure of increasing K from 1 to 6. Each step resulted in a significant cluster separation leading us to the final choice of six clusters. When seven clusters were imposed we still find the former six clusters and an extra “latent” cluster that in the final year-by-year cluster assignment by majority vote never comes out as the modal one, leaving the results unchanged and supporting six as a suitable number of clusters. To choose the optimal length of Voronoi elements, we use the entropy criterion described in

Sect. 3.3. In Fig. 7 we report the annual bootstrap entropies for $L = 30, 100, 200$ and 400. We see that $L = 200$ provides the minimal average entropy and thus motivates our choice for this parameter for the analysis.

5 A simulation study

To further exemplify our method we present the following simulated data example, representing dynamics of yearly weather profiles along 6000 years. The dynamics are modeled with a two layer model, a hidden layer corresponding to climate and an observable layer corresponding to weather. The hidden climate layer is modeled using a two state Markov chain with transition matrix

$$\mathbb{P} = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}.$$

This corresponds to periods of distinct climates of length of 100 years on average. The simulated climate dynamics are illustrated in the upper part of Fig. 9. There are two different types of weather. The frequencies of the two weather types are different in the two climatic states, occurring with probabilities $[0.7, 0.3]$ for each year in climate state 1, and with probability $[0.3, 0.7]$ in climate state 2. The

Fig. 6 Boxplots of reconstructed Stockholm winter temperatures (January–April) 1502–1892, Leijonhufvud et al. (2010) for the 6 KMA clusters. The corresponding winter temperatures (January–April) in the Kassjön area is about 5 degrees lower. The number of years falling in each cluster is given above the boxplots (missing values excluded)

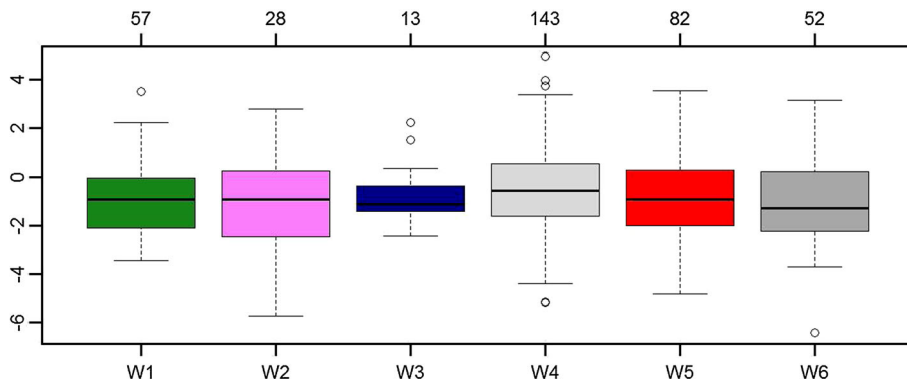
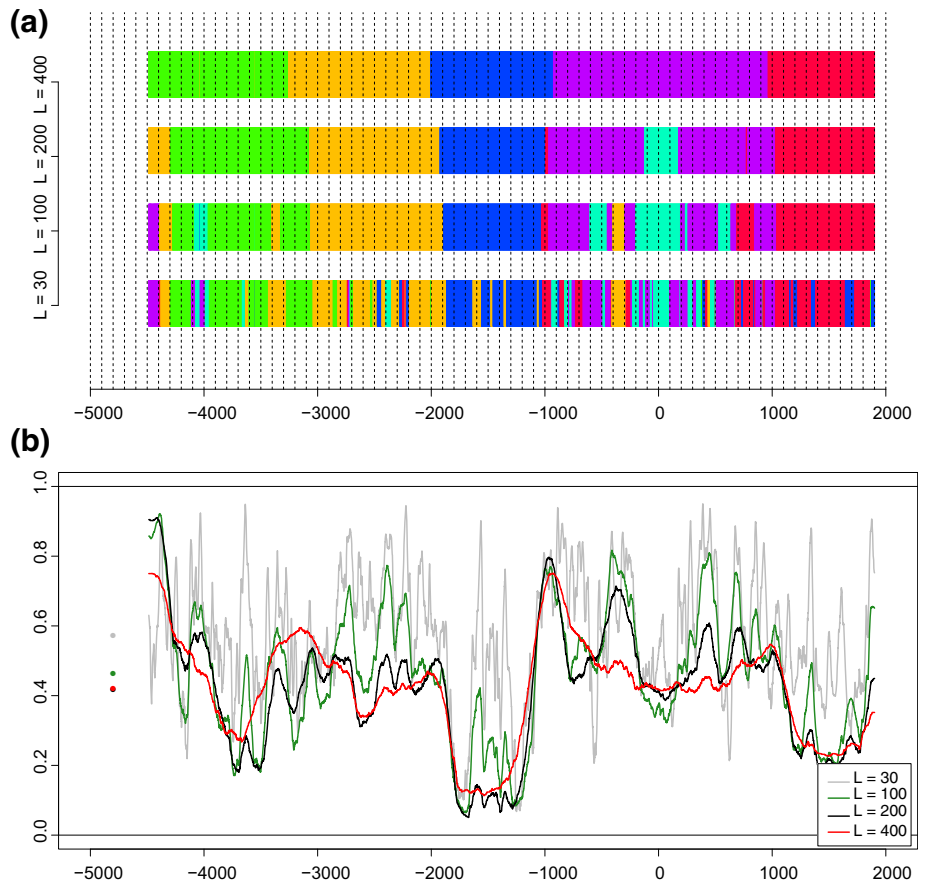


Fig. 7 a The BVKMA cluster dynamics for $K = 6$ clusters when $L = 30, 100, 200$ and 400 . **b** The entropies of the relative frequencies of assignment for each year for the different L 's. The dots to the left correspond to the average (over years) entropies



observable weather layer is constructed based on the two functional weather templates

$$w_i(t) = c_{1i} \sin(2\pi t) + c_{2i} \sin(4\pi t) + c_{3i} \cos(4\pi t), \quad i = 1, 2,$$

observed on the interval $[0,1]$, with $c_{11} = -1$, $c_{21} = 1$, $c_{31} = 2$ and $c_{12} = 0$, $c_{22} = 0$, $c_{32} = 2$, illustrated in Fig. 8(a) on the interval $[0,1]$.

The observed curves are not exact observations of the weather templates. They are affected by amplitude and phase variation. The variation in amplitude is exhibited by mean zero independent normally distributed variation around each of the coefficients c_{ij} with standard deviation $\sigma = 0.05$, i.e. $\mathcal{N}(0, 0.05^2)$. Random variation in phase is incorporated by introducing independent random warping functions, $h(t) = a + bt$ with $a \sim \mathcal{N}(0, 0.01^2)$ and $b \sim \mathcal{N}(1, 0.05^2)$, for each year. An example of simulated weather curves is visualised in Fig. 8(c). We further introduce an additional random phase variation that applies simultaneously to all weather curves within the same climate period (i.e., along the consecutive sequence of years with the same climate label): For all weather curves within each climate period, an additional phase shift of size $1/4\pi$ is applied with probability $1/2$. The weather templates shifted by $1/4\pi$ are illustrated by dashed lines in Fig. 8(b).

Incorporating all random variation in amplitude and phase, we thus construct the observed weather curves used in our analysis and presented in Fig. 8(d). Two different amplitude formations, distorted or not by phase warping are visible in Fig. 8(d). We stress that Fig. 8(d) only presents the observable weather forms, not the underlying climate. It is the frequency of weather types within longer time periods that constitutes climate.

We now analyze the data by the introduced clustering methods, addressing different aspects of the observed data. In line with the analysis of the Kassjón data, the methods (when applicable) use the normalized L^2 distance to measure the distance between functions and the family of affine warping functions. We first run the KMA algorithm in an attempt to recover weather, taking into account the misalignment. We used the within cluster variation argument of Sangalli et al. (2010b) to determine the number of clusters, resulting in the (correct) choice of $K = 2$ weather clusters. For this choice of $K = 2$ the KMA in this case correctly recovers 99 % of the weather labels and 69 % of the climate labels. Hence, the KMA method almost perfectly recovers weather, but not the underlying climate. As a comparison, the functional K-Medoid algorithm with $K = 2$, that is not taking misalignment (or spatial

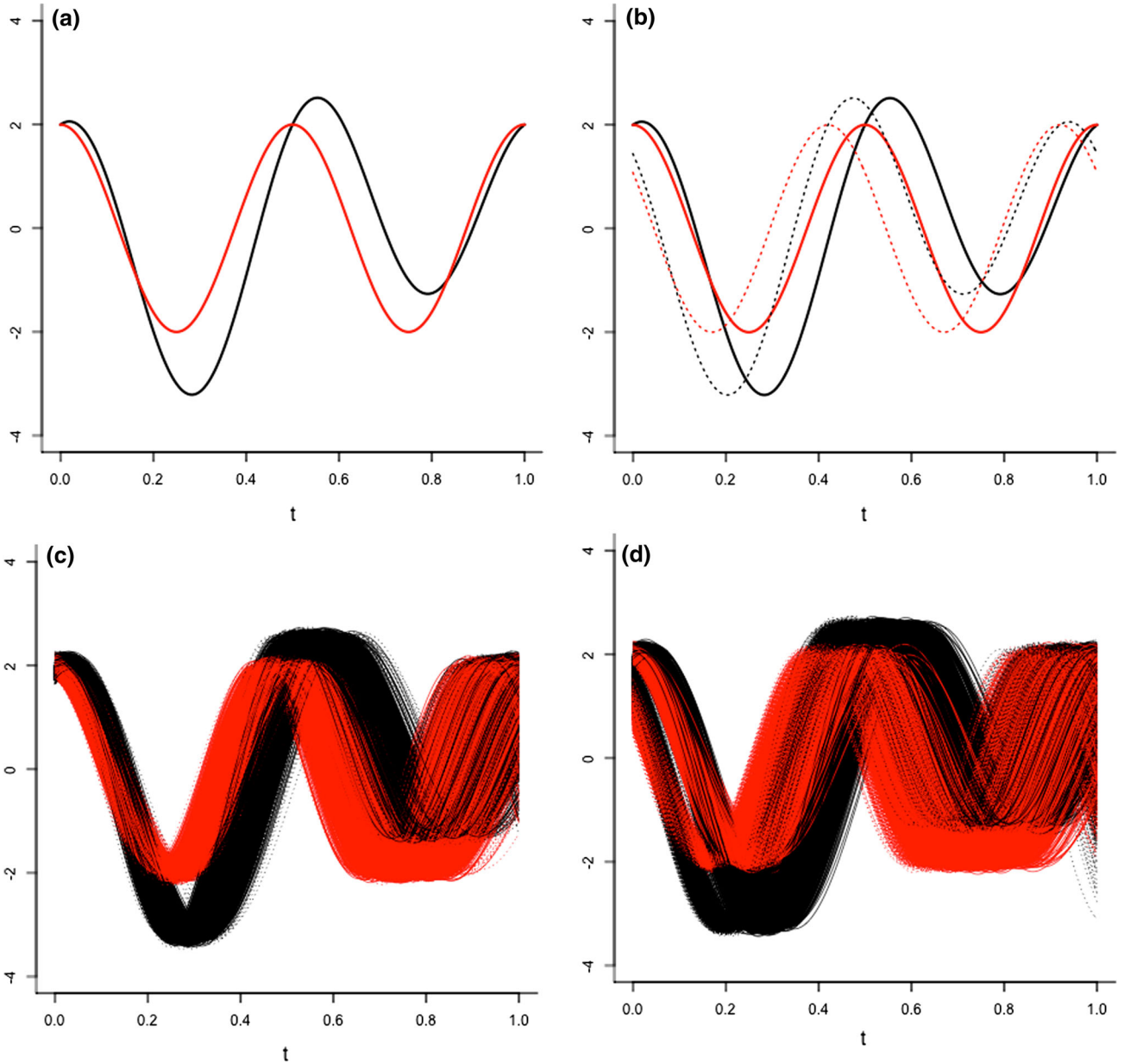


Fig. 8 **a** The weather templates $w_1(t)$ (black line) and $w_2(t)$ (red line). **b** The weather templates $w_1(t)$ and $w_2(t)$ (solid lines) and corresponding templates shifted by $1/4\pi$ (dashed lines). Simulated (misaligned) weather curves, in **(c, d)**, based on the templates in **(a, b)**, respectively

dependence) into account, correctly recovers only 60 % of the weather labels and 55 % of the climate labels.

We started the analysis with BVKMA by determining the number of clusters K to use by running the algorithm with $B = 500$ and $L = 50$ while varying number of clusters K . For $K = 2$, we obtained a clear separation of the clusters and both clusters were shown by the majority vote. For $K = 3$ we also obtained 3 separate clusters with distinct majority votes in the aggregation phase. Investigating the structure of those 3 clusters it was noted that the first cluster was the same as one of the clusters in the case of

$K = 2$. The second and third clusters emerged from a separation of the second cluster from the case of $K = 2$. Moreover, the centroids for the last two clusters were almost identical to each other with respect to amplitude, differing only by phase shift. A further increase of K to 4 resulted in the emerging of a latent cluster which was never selected in the final classification by majority vote. This led us to decide upon $K = 2$, which also is the correct number of climate clusters.

The second step of the BVKMA method calibration is to decide the average length L of the Voronoi elements, L . We

used the minimum average entropy criterion discussed in Sect. 3.3 for this purpose. For $K = 2$ we calculated the average entropy for a raster of values of L (see, Table 1) and finally chose $L = 25$ which corresponded to the smallest average entropy. The BVKMA with $K = 2$ and $L = 25$ turned out to correctly recover 94.52% of the climate labels. When considering the proportions of weather types (detected with KMA) arising within the two different climate clusters we obtained [0.66 0.34] and [0.30 0.70], respectively, in line with the expected true proportions [0.7 0.3] and [0.3 0.7]. Hence not only the climate labels but also the corresponding within cluster structure was correctly recovered. The time dynamics of the climate labels are presented in Fig. 9, the upper segment presenting the true dynamics of the simulated climate states and the lower segment presenting the recovered labelling obtained with the BVKMA algorithm.

Table 1 summarizes the correctly recovered weather and climate labels by the BVKMA algorithm for $K = 2$ and various values of L . Note that the special case of BVKMA with $L = 1$ corresponds to the KMA algorithm (with $K = 2$). The selection of L suggested by the entropy criterion is almost optimal compared to other choices.

To emphasize the importance of taking the phase variation into consideration, we compare the results of BVKMA to those for the BVKM method with $K = 2$ and various average lengths L of the Voronoi elements. The highest recovery climate label rate attained for BVKM was 64.47% with $L = 50$. Summarizing, we see that in order to correctly recover the climate labels we needed to take into account the local dependency structure together with the misalignment. Omitting any of these two factors led to significantly lower climate recovery rates, none of them exceeding 70%.

6 Discussion and concluding remarks

We propose a new functional clustering method, the Bagging Voronoi K-Medoid Alignment (BVKMA) algorithm, which to our knowledge is the first method that jointly handles dependent and misaligned functions. It has been obtained by suitably merging a clustering technique for dependent functional data (Bagging Voronoi K-Medoid) and a clustering technique for misaligned functional data (K-Medoid Alignment). The method is general and can

deal, in a non-parametric fashion, with various dependency structures and possibly also different clustering techniques. It is flexible and can be adapted to arbitrary families of warping functions allowing for adjustments for any type of misalignment. Additionally, the method is not limited to one-dimensional settings and can be straightforwardly applied to higher dimensional problems including spatial and spatio-temporal dependency. The simulation study exemplifies the importance and superiority of using clustering methods that jointly handle the dependence and misalignment of the curves, such as the BVKMA method, in recovering the correct cluster labels.

When applied to the Kassjön sediment data, the method provides a way to summarize the weather variability in terms of longer term changes on different time scales, corresponding to climate. We detected six different climate regimes aiming to capture climate. They are all characterized by significantly different frequencies of seasonal pattern (weather) types detected by the K-Medoid algorithm. Two of the climate periods, (4300 BC, 3100 BC) and (150 BC, AD 150), have high frequencies of years with pronounced spring peak greyscale patterns, indicating an intense spring flood and high snow accumulation during winter. Climate periods (1950 BC, 1000 BC) and (AD 1000, AD 1900), on the other hand, are characterized by high frequencies of years with flatter seasonal greyscale profiles, indicating less winter (snow) precipitation and milder winters. Years with significant sediment accumulation after the spring flood are frequent in the climate regime during (3100 BC, 1950 BC), perhaps indicating warmer summers and/or fall storms. For climate period (1000 BC, AD 1000) excluding (150 BC, AD 150) all different weather types are approximately equally likely.

Confirmation of the climatic interpretations by a validation of the varve characteristics by means of regional meteorological observations is unfortunately difficult due to potential dating uncertainties and human impact (agricultural activities and ditching) during the last centuries. Some preliminary comparisons with other regional climate archives still seem to support the climate interpretations of the clusters. However, a more thorough investigation, beyond the scope of this paper, is necessary to confirm the climatic findings, even though the novel approach presented here shows great potential of revealing past climate.

Table 1 The mean entropies together with weather and climate label recovery rates for BVKMA analyses with $K = 2$ and various L . For each quantity the best achieved outcome is reported in bold

L	1 (KMA)	5	10	25	50	100	200
Mean entropy	–	0.56	0.39	0.30	0.33	0.39	0.44
Climate recovery	69 %	89.6 %	94.92 %	94.52 %	90.77 %	86.38 %	81.85 %
Weather recovery	99 %	75.03 %	70.32 %	68.75 %	66.80 %	65.05 %	62.88 %

Recovery rates for $L = 1$ corresponds to KMA analyses with $K = 2$

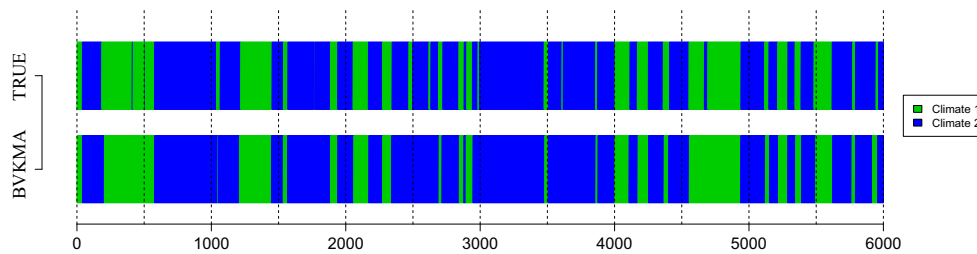


Fig. 9 True and recovered climate dynamics in the simulated dataset. The recovered labels are obtained with help of BVKMA algorithm with $K = 2$ and $L = 25$

Acknowledgments The authors would like to thank Christian Bigler for valuable comments and discussions. This work was supported by the Swedish Research Council, project D0520301.

References

- Arnqvist P, Bigler C, Renberg I, Sjöstedt de Luna S (2016) Functional clustering of varved lake sediment to reconstruct past seasonal climate. *J Environ Ecol Stat* (351). doi:10.1007/s10651-016-0351-1
- Beniston M (2005) Warm winter spells in the Swiss Alps: strong heat waves in a cold season? A study focusing on climate observations at the Saentis high mountain site. *Geophys Res Lett* 32(1)
- Comas C, Mehtätalo L, Miina J (2013) Analysing spacetime tree interdependencies based on individual tree growth functions. *Stoch Environ Res Risk Assess* 27(7):1673–1681
- Dabo-Niang S, Yao AF, Pischedda L, Cuny P, Gilbert F (2010) Spatial mode estimation for functional random fields with application to bioturbation problem. *Stoch Environ Res Risk Assess* 24(4):487–497
- Fernández-Pascual R, Espejo R, Ruiz-Medina M (2015) Moment and bayesian wavelet regression from spatially correlated functional data. *Stoch Environ Res Risk Assess* 30:523–557
- Finazzi F, Haggarty R, Miller C, Scott M, Fasso A (2015) A comparison of clustering approaches for the study of the temporal coherence of multiple time series. *Stoch Environ Res Risk Assess* 29:463–475
- Gaffney SJ, Smyth P (2004) Joint probabilistic curve clustering and alignment. In: *Advances in neural information processing systems*, pp 473–480
- Giraldo R, Delicado P, Mateu J (2012) Hierarchical clustering of spatially correlated functional data. *Stat Neerl* 66(4):403–421
- Ignaccolo R, Ghigo S, Giovenali E (2008) Analysis of air quality monitoring networks by functional clustering. *Environmetrics* 19(7):672–686
- Ignaccolo R, Mateu J, Giraldo R (2014) Kriging with external drift for functional data for air quality monitoring. *Stoch Environ Res Risk Assess* 28(5):1171–1186
- Leijonhufvud L, Wilson R, Moberg A, Söderberg J, Retsö D, Söderlind U (2010) Five centuries of Stockholm winter/spring temperatures reconstructed from documentary evidence and instrumental observations. *Clim Change* 101(1–2):109–141
- Liu X, Müller HG (2004) Functional convex averaging and synchronization for time-warped random curves. *J Am Stat Assoc* 99(467):687–699
- Mann ME, Zhang Z, Hughes MK, Bradley RS, Miller SK, Rutherford S, Ni F (2008) Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc Natl Acad Sci USA* 105(36):13252–13257
- Menafoglio A, Secchi P, Guadagnini A (2016) A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. *Math Geosci* 48(4):463–485
- Menafoglio A, Guadagnini A, Secchi P (2014) A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stoch Environ Res Risk Assess* 28(7):1835–1851
- Ojala AE, Alenius T (2005) 10000 years of interannual sedimentation recorded in the Lake Nautajärvi (Finland) clastic-organic varves. *Palaeogeogr, Palaeoclimatol, Palaeoecol* 219(3):285–302
- Ojala AE, Alenius T, Seppä H, Giesecke T (2008) Integrated varve and pollen-based temperature reconstruction from Finland: evidence for Holocene seasonal temperature patterns at high latitudes. *Holocene* 18(4):529–538
- Pachauri RK, Allen MR, Barros VR, Broome J, Cramer W, Christ R, Church JA, Clarke L, Dahe Q, Dasgupta P, Dubash NK, Edenhofer O, Elgizouli I, Field CB, Forster P, Friedlingstein P, Fuglestvedt J, Gomez-Echeverri L, Hallegatte S, Hegerl G, Howden M, Jiang K, Cissneroz BJ, Kattsov V, Lee H, Mach KJ, Marotzke J, Mastrandrea MD, Meyer L, Minx J, Mulugetta Y, O’Brien K, Oppenheimer M, Pereira JJ, Pichs-Madruga R, Plattner GK, Pörtner HO, Power SB, Preston B, Ravindranath NH, Reisinger A, Riahi K, Rusticucci M, Scholes R, Seyboth K, Sokona Y, Stavins R, Stocker TF, Tschakert P, van Vuuren D, van Ypserle JP (2014) Climate change 2014: synthesis report. contribution of working groups I, II and III to the 5th assessment report of the intergovernmental panel on Climate Change. IPCC, Geneva, Switzerland
- Petterson G (1999) Image analysis, varved lake sediments and climate reconstruction. Ph.D thesis, Umeå University
- Petterson G, Renberg I, Geladi P, Lindberg A, Lindgren F (1993) Spatial uniformity of sediment accumulation in varved lake sediments in northern Sweden. *J Paleolimnol* 9(3):195–208
- Petterson G, Odgaard B, Renberg I (1999) Image analysis as a method to quantify sediment components. *J Paleolimnol* 22(4):443–455
- Petterson G, Renberg I, Sjöstedt-de Luna S, Arnqvist P, Anderson NJ (2010) Climatic influence on the inter-annual variability of late-Holocene minerogenic sediment supply in a boreal forest catchment. *Earth Surf Process Landf* 35(4):390–398
- R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org/>
- Ramsay J, Silverman B (2005) Functional data analysis. Springer, New York
- Romano E, Balzanella A, Verde R (2010) Classification as a tool for research. In: *Proceedings of the 11th IFCS Biennial conference and 33rd annual conference of the Gesellschaft für Klassifikation e.V.*, Dresden, March 13–18, 2009, Springer, Heidelberg, chap clustering spatio-functional data: a model based approach, pp 167–175

- Romano E, Mateu J, Giraldo R (2015) On the performance of two clustering methods for spatial functional data. *AStA Adv Stat Anal* 99(4):467–492
- Salazar E, Giraldo R, Porcu E (2015) Spatial prediction for infinite-dimensional compositional data. *Stoch Environ Res Risk Assess* 29(7):1737–1749
- Sangalli LM, Secchi P, Vantini S, Vitelli V (2010a) Functional clustering and alignment methods with applications. *Commun Appl Ind Math* 1(1):205–224
- Sangalli LM, Secchi P, Vantini S, Vitelli V (2010b) K-mean alignment for curve clustering. *Comput Stat Data Anal* 54(5):1219–1233
- Sangalli LM, Secchi P, Vantini S (2014) Analysis of aneurisk65 data: *k*-mean alignment. *Electron J Stat* 8(2):1891–1904
- Secchi P, Vantini S, Vitelli V (2011) Spatial clustering of functional data. In: *Recent advances in functional data analysis and related topics*, Springer, New York, pp 283–289
- Secchi P, Vantini S, Vitelli V (2013) Bagging Voronoi classifiers for clustering spatial functional data. *Int J Appl Earth Obs Geoinf* 22:53–64
- Seegerström U, Renberg I, Wallin JE (1984) Annual sediment accumulation and land use history; investigations of varved lake sediments. *Verh int Ver Limnol* 22:1396–1403
- Stephens M (2000) Dealing with label switching in mixture models. *J R Stat Soc* 62(4):795–809
- Tarpey T, Kinader KK (2003) Clustering functional data. *J Classif* 20(1):93–114
- Tiljander M, Saarnisto M, Ojala AE, Saarinen T (2003) A 3000-year palaeoenvironmental record from annually laminated sediment of Lake Korttajarvi, central Finland. *Boreas* 32(4):566–577
- Vantini S (2012) On the definition of phase and amplitude variability in functional data analysis. *Test* 21(4):676–696