Received: 15 June 2020 | Revised: 6 May 2021 | Accepted: 23 June 2021

DOI: 10.1002/sim.9130

RESEARCH ARTICLE

Statistics in Medicine WILEY

Continuous positive airway pressure adherence trajectories in sleep apnea: Clustering with summed discrete Fréchet and dynamic time warping dissimilarities

Guillaume Bottaz-Bosson[1,2] | Agnès Hamon[2] | Jean-Louis Pépin[1] | Sébastien Bailly[1] | Adeline Samson[2]

[1] Laboratoire HP2, Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, Grenoble, France

[2] LJK, Univ. Grenoble Alpes, CNRS, Grenoble, France

Correspondence
Guillaume Bottaz-Bosson, Laboratoire HP2, Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, 38000 Grenoble, France.
Email: guillaume.bottaz-bosson@univ-grenoble-alpes.fr

Background: Obstructive sleep apnea (OSA) is a chronic disease characterized by recurrent pharyngeal collapses during sleep. In most severe cases, continuous positive airway pressure (CPAP) consists in keeping the airways open by administering mild air pressure. This treatment faces adherence issues.

Objectives: Eight hundred and forty-eight subjects were equipped with CPAP prescribed at the Grenoble University Hospital between 2016 and 2018. Their daily CPAP uses have been recorded during the first 3 months. Our aim is to cluster these adherence time series. With hierarchical agglomerative clustering, we focused on the choices of the dissimilarity measure and the internal cluster validation index (CVI).

Methods: The Euclidean distance, the dynamic time warping (DTW) and the generalized summed discrete Fréchet dissimilarity were implemented with three linkage strategies ("average," "complete," and "Ward"). The performances of each method (dissimilarity and linkage) were evaluated on a simulation study through the adjusted Rand index (ARI). The Ward linkage with DTW dissimilarity provided the best ARI. Then six different internal CVIs (Silhouette, Calinski Harabasz, Davies Bouldin, Modified Davies Bouldin, Dunn, and COP) were compared on their ability to choose the best number of clusters. The Dunn index beat the others.

Results: CPAP data were clustered with the Ward linkage, the DTW dissimilarity and the Dunn index. It identified six clusters, from a cluster of patients (N = 29 subjects) whose stopped the therapy early on to a cluster (N = 105) with increasing adherence over time. Other clusters were extremely good users (N = 151), good users (N = 150), moderate users (N = 235), and poor adherers (N = 178).

KEYWORDS

cluster validation, CPAP adherence trajectories, discrete Fréchet distance, dynamic time warping, time series clustering

1 | INTRODUCTION

1.1 | Clinical context and rationale

Obstructive sleep apnea (OSA) is one of the most common chronic diseases affecting almost 1 billion people worldwide.[1] OSA is characterized by repeated episodes of complete (apneas) and/or incomplete (hypopneas) pharyngeal collapse during sleep producing intermittent hypoxia and sleep fragmentation which in turn generate disturbing symptoms including daytime sleepiness, impairment of daily functioning, deterioration of memory and cognition, and a higher risk of developing cardiovascular, metabolic, and cerebrovascular disease.[2]

Continuous positive airway pressure (CPAP) is the first line treatment of moderate to severe OSA. CPAP reopens and stabilizes the upper airway and allows the complete suppression of abnormal respiratory events during sleep. The therapy is highly effective in improving quality of life and suppressing symptoms, but adherence remains challenging. A period of 3 to 6 months after CPAP initiation is necessary for some patients to stabilize their adherence due to the need to get used to the mask interface, to the pressure, to using the CPAP device, and adapt to the side effects. The initial refusal rate by patients is close to 15% even when proposed by experienced teams, and long term treatment discontinuation is estimated at between 20% and 35%.[2] The majority of dropouts happens in the first 3 months.[3] While variable from one country to another, a minimal adherence to CPAP is required for CPAP treatment reimbursement (eg, in the United States more than 4 hours of usage per night for more than 70% of nights). There is a dose response relationship between CPAP adherence and the degree of improvement in symptoms and related quality of life.[4] The potential for cardio-metabolic risk reduction is also highly dependent on CPAP adherence levels.[2]

A unique specificity of OSA is that data are generated every night by telemonitoring of CPAP devices in millions of patients providing objective daily measurements of adherence.[5] Interventions to improve CPAP adherence have included educational, supportive, and behavioral strategies[6] or technical CPAP innovations to reduce device-related side effects. When implemented separately, these approaches have only had a limited impact on CPAP adherence[6] and recent strategies have aimed at combining information from remote home monitoring of CPAP use and patient coaching. In depth detailed analysis of CPAP telemonitoring data is a crucial step toward describing the different patterns of CPAP adherence. This is a prerequisite to identifying patients at risk of poor CPAP adherence and to proposing personalized follow-up adapted to these profiles. The main goal of the present work is to propose statistical tools to delineate the different trajectories of CPAP adherence during the first three months of follow-up.

1.2 | Methodological approaches

Every night CPAP adherence data available from telemonitoring allow us to construct chronological diagrams of individualized trajectories of adherence. Here we employ a time series clustering approach, to describe the typical patterns and reveal the most illustrative CPAP adherence trajectories. Clustering algorithms are generally designed to deal with static data and therefore are not completely appropriate to consider the temporal structure of time series. Transforming time series into static data (by estimating model parameters or feature extraction) is one method. Customizing the algorithms is another. Moreover there are various clustering algorithms that can produce different results from the same data. Furthermore, a key issue is to evaluate the goodness of the resulting partition as there is no prior information about the data. Consequently, it is difficult to both choose an appropriate clustering method and to select the number of clusters in which to split the data. A first distinction among time series clustering methods is between model-based and shape-based approaches. Model-based clustering assumes a specific model for each cluster, and suitable model distance measures and algorithms are applied. The main drawback is that the results rely on the model's assumptions. Conversely, shape-based methods do not require any assumptions and work on raw data by using an appropriate dissimilarity measure. In the present study, we preferred a shape-based approach because of the small amount of prior knowledge available on CPAP adherence trajectories. In these shape-based approaches, the two main unsupervised classification algorithms are partitional and hierarchical clustering. Both rely on the choice of a dissimilarity measure. The Euclidean distance is frequently used with these algorithms but is not specific to the context of time series clustering, contrary to the dynamic time warping (DTW) dissimilarity which considers shapes differing only in temporal shift as being similar. Along the same line, Genolini et al[7] introduced the generalized discrete Fréchet distance that comports a time scale parameter to take account of the time shifts in the dissimilarity value. We propose a variant of this dissimilarity called the "generalized summed discrete Fréchet dissimilarity" (sdF) which is a generalization of the DTW including a time scale parameter. In this context, we consider both DTW and sdF dissimilarities because time shifts are not important from the medical point of view. For example, when we want to consider all patients who abandon their CPAP therapy as being in the same group, whatever the duration between CPAP initiation and treatment surrender. We are interested in how these dissimilarity measures identify shapes despite the particularities of CPAP trajectories, from recurrent discontinuities due to the random occurrence of zero use to high variability hiding central tendency. These particularities are of interest when describing adherence behaviors and they motivate the choice to cluster the data without applying any smoothing method. However, the relevance of the dissimilarities are dependent on the clustering algorithm used. Partitional clustering requires the choice of a centroid algorithm and setting the required number of clusters (k). The initialization step randomly spreads the data into k clusters and computes centroids, called cluster centers. Then the algorithm alternates between reallocating the objects to the clusters with the nearest cluster centers and computation of the new cluster centers (centroids) until a quite stable partition is obtained. Two notable variants of this algorithm are K-medoids and K-means. K-medoids work with arbitrary dissimilarities and use medoids as centroids. K-means are clustering using the Euclidean distance and the centroids are defined as the arithmetic mean between the points in a cluster. DTW and sdF dissimilarities can be used in partitional clustering with suitable centroids (see Section 3.4), but involve high computation cost, and a new calculation is required at each iteration, so the overall computational cost of the algorithm is strongly impacted. Furthermore, the resulting clusters are dependent on the initialization step and so in practice, for a chosen k, several clusterings are done and compared before validating the final partition. Considering the necessity to launch several executions, for both initialization reasons and the k-parameter input, we excluded partitioning algorithms from this study so as to concentrate on hierarchical approaches. Thus we focus on hierarchical agglomerative clustering (HAC): after the computation of a pairwise dissimilarity matrix, each object is placed in an individual cluster which are then gradually merged through a linkage strategy. Then a dendrogram presents the hierarchy of clusters and the resulting partition is obtained after cutting it at the desired height or with the adopted number of clusters. Optionally, centroids are only computed once during this final step.

In the case of time series clustering, as for classical clustering, the evaluation of the resulting clusters and the choice of cluster number remain open questions. Many cluster validity indices (CVIs) exist in the literature. They are split into two categories: external and internal indices. The first ones evaluate the similarity between two partitions of a same set. They judge the quality of a partition when, for example, the true partition is known. On the other hand, internal indices can measure the goodness of a partition, without external information. Each of them gives a different sense of what is a "good" partition. Gurrutxaga et al[8] proposed a methodology to compare internal CVIs. One novelty of our work is to transpose this approach to the case of time series clustering, which has not been done yet.

To recap, we focused on HAC where we needed first to select a dissimilarity measure and then to validate a CVI to choose the number of clusters. These two choices motivated the current study and this article is organized as follows. In the next section, we present the data (Section 2). Section 3 describes several alternatives we considered in the framework of the approach described above, and they were compared in a simulation study (Section 4). Then, in Section 5, we describe the clusters resulting after application of the selected clustering procedure to real data. Section 6 concludes the article with some discussion and potential applications of these method to time series produced in other fields of medicine.

2 | DATA

2.1 | Confusion between missing and null values

A problem with CPAP adherence data is the possible confusion between missing and null values. Missing data is the absence of an adherence value whereas a null value means the patient did not use her/his device. As a missing value can be due to a technical problem with the CPAP machine, or due to the fact a patient did not power her/his device, it is difficult to differentiate null and missing values. Different machines do not manage these situations in the same way so these values must be interpreted for each different CPAP model.

2.2 | Population selection

Our data included N = 1831 subjects who started a CPAP therapy between January 2016 and January 2018 prescribed by the Grenoble-Alpes University Hospital (France). The adherence follow-up was launched within 15 days after treatment

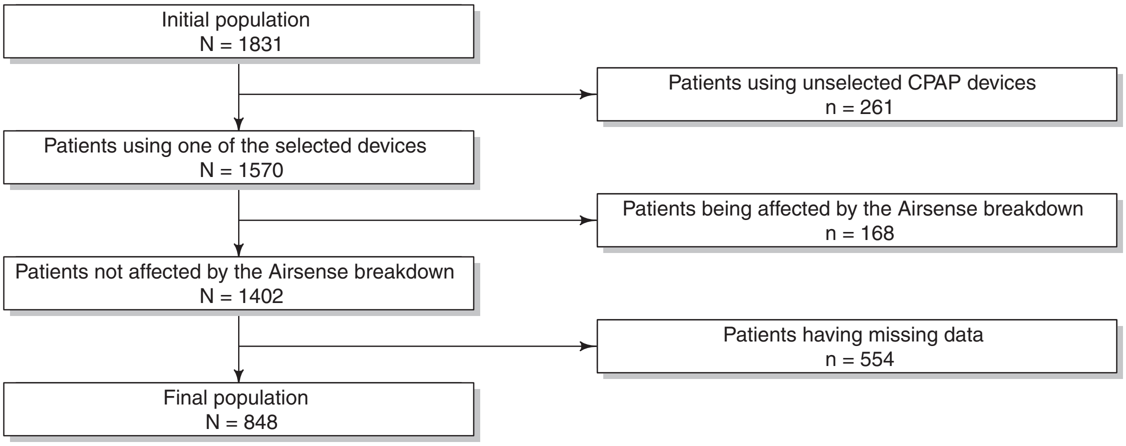

Initial population
N = 1831
Patients using unselected CPAP devices
n = 261
Patients using one of the selected devices
N = 1570
Patients being affected by the Airsense breakdown
n = 168
Patients not affected by the Airsense breakdown
N = 1402
Patients having missing data
n = 554
Final population
N = 848

FIGURE 1 Real data: Subject inclusion flowchart

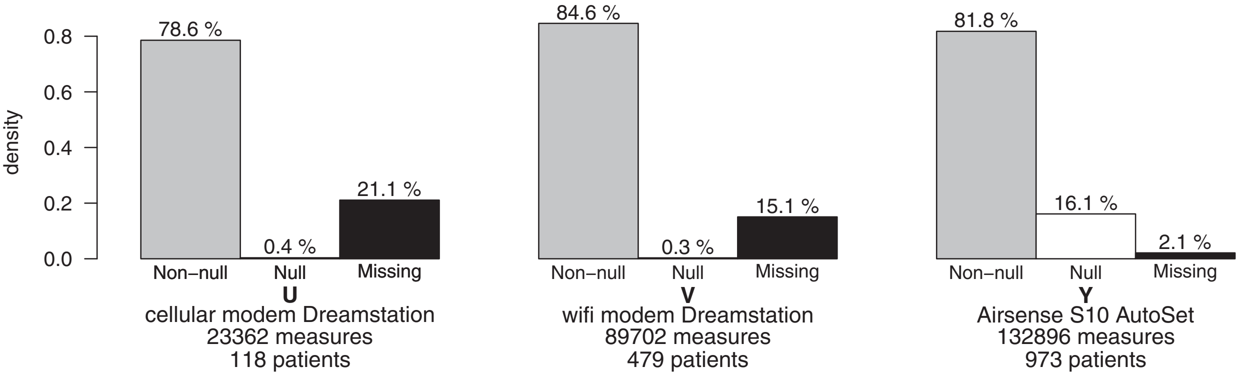


FIGURE 2 Real CPAP data: Distribution of adherence values (non-null, null, and missing) for the 3 selected CPAP devices

prescription. Figure 1 resumes the flowchart of patient inclusion in the study. These subjects did not change their CPAP device and do not have days with multiple measures during the period between 2016 and 2018.

A step preceding selection of the population to be analyzed was to distinguish missing from null values for each CPAP model. Therefore we considered only models of CPAP devices used by at least 100 subjects and we excluded CPAP devices that are unable to transmit data without being connected to an external modem. Indeed, for the latter, the data corresponds to modem utilization, which can be used without the CPAP device or inversely the device can be used without the modem. Three devices used by 1570 subjects were retained. We call these devices U, V, and Y where U and V are Dreamstation devices (Philips) with respectively internal cellular modems or dependent on WiFi, and Y is the Airsense S10 Autoset device (ResMed). Figure 2 shows the distribution of missing, null, and non-null adherence data, for the three selected devices, confirming that the meaning of “missing” and “null” differs according to the device. For machines U and V, and in view of the low zero rates, we considered that if a subject did not use her/his CPAP, the device remained off and so there was no data transmitted. Missing data for these devices were then replaced by null adherence. For the device Y, data were reliable as the modem sent a “zero” even if the patient did not use her/his device.

To simplify the statistical analysis, we considered trajectories without missing values and the length of the time series was fixed at 91 days. A breakdown occurred with the Y device between June 9, 2017 and June 18, 2017. This affected the transmission of adherence data of 168 subjects, who we excluded. Finally we excluded 554 more subjects with missing values, to obtain a final population of 848 patients with complete trajectories.

Table 3 briefly describes this population. They were middle aged (75% over 50 years), predominantly male (63%) and with a median body mass index of 30.4 kg/m^2. This table also summarizes individual adherence characteristics with individual means, standard deviations, rates of null values, rates of use for over 4 hours. The distributions of these characteristics are shown in the histograms in Appendix A (Figure A1). Figure 3 shows 8 examples of individual

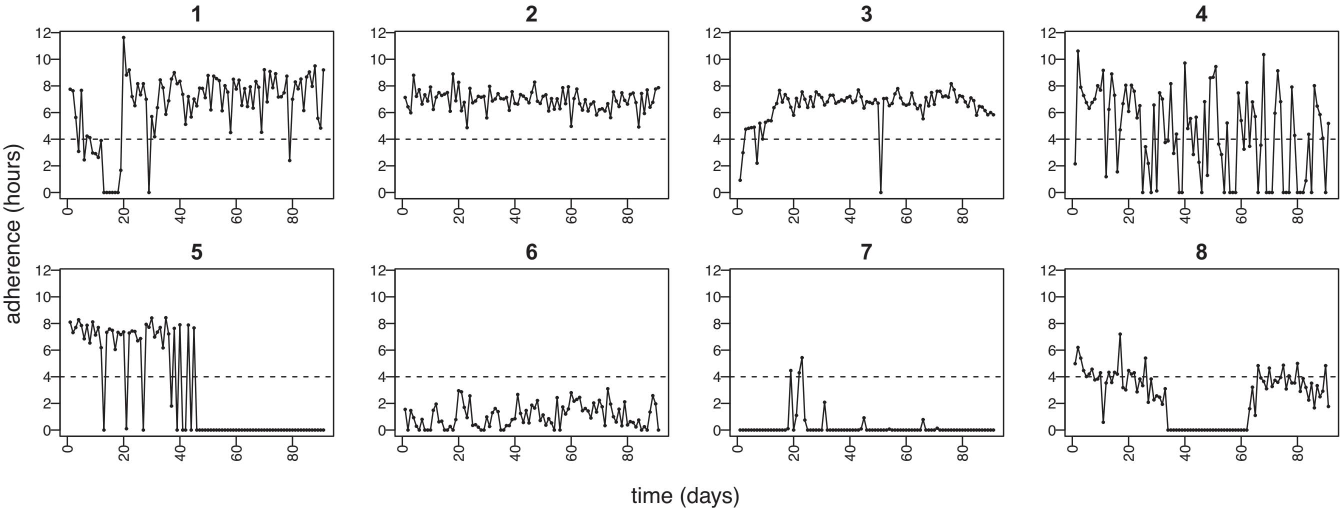


FIGURE 3 CPAP adherence time series during the first 3 months of therapy of 8 patients. The dashed line represents the efficiency threshold fixed at 4 hours of daily CPAP use

adherence trajectories, including subjects having good adherence trajectories (subject 2) patients with occasional or insufficient CPAP use (subjects 6 and 7) and patients who completely stopped CPAP (subject 5). High variability and random occurrences of zeros characterize these CPAP time series.

We remind readers that the clinical objective of this work was to find typical CPAP adherence profiles by clustering the data. This is a first step towards improving the personalized care of patients with sleep apnea. Several alternatives methods to cluster CPAP data in the framework of HAC are detailed in the next section.

3 | CLUSTERING PROCEDURE

Clustering time series using the HAC algorithm involves the computation of a pairwise dissimilarity matrix, a linkage strategy to define distances between clusters and the choice of the number of clusters. This choice can be made after inspecting the dendrogram node heights, but also thanks to internal CVIs. We focused on this second approach because it is more objective, although sometimes it can be dependent on the calculation of cluster centroids.

3.1 | Some dissimilarity measures between time series

Here, we describe the three dissimilarity measures considered in this study. Formal mathematical definitions are given in Appendix B.2.

The Euclidean distance computes the pointwise difference between the two time series. It is the square root of the summed daily squared differences (see Definition 3). Figure 4 illustrates the three dissimilarity measures on two fictive trajectories. The two trajectories have the same pattern with a time shift of one unit and differing from one unit on the y axis. This measure is fast to compute and compatible with the arithmetic mean used as a centroïd, making it easy to implement in all algorithms. However it is not specific to time series, contrary to DTW for example (see Definition 4).

This last measure is frequently used in time series clustering.[9] It is said to be “elastic” because the time series are warped and two patterns differing only in a time shift are thus considered as similar. The main drawbacks of this dissimilarity are the calculation cost, higher than for the Euclidean distance; and the choice of the centroid, as the arithmetic mean of several time series sometimes produces a series with a shape that is very different from all the individual series. We note that this measure considers patients stopping therapy as close and would identify a “dropout” cluster.

The Fréchet distance[10] applies to parametric curves. A discrete version exists for polygonal curves and can be applied to time series through the *generalized discrete Fréchet distance* and its variants.[7] We propose a formal definition of the generalized Fréchet distance and its variant, the *generalized summed discrete Fréchet dissimilarity* (Definition 5). These

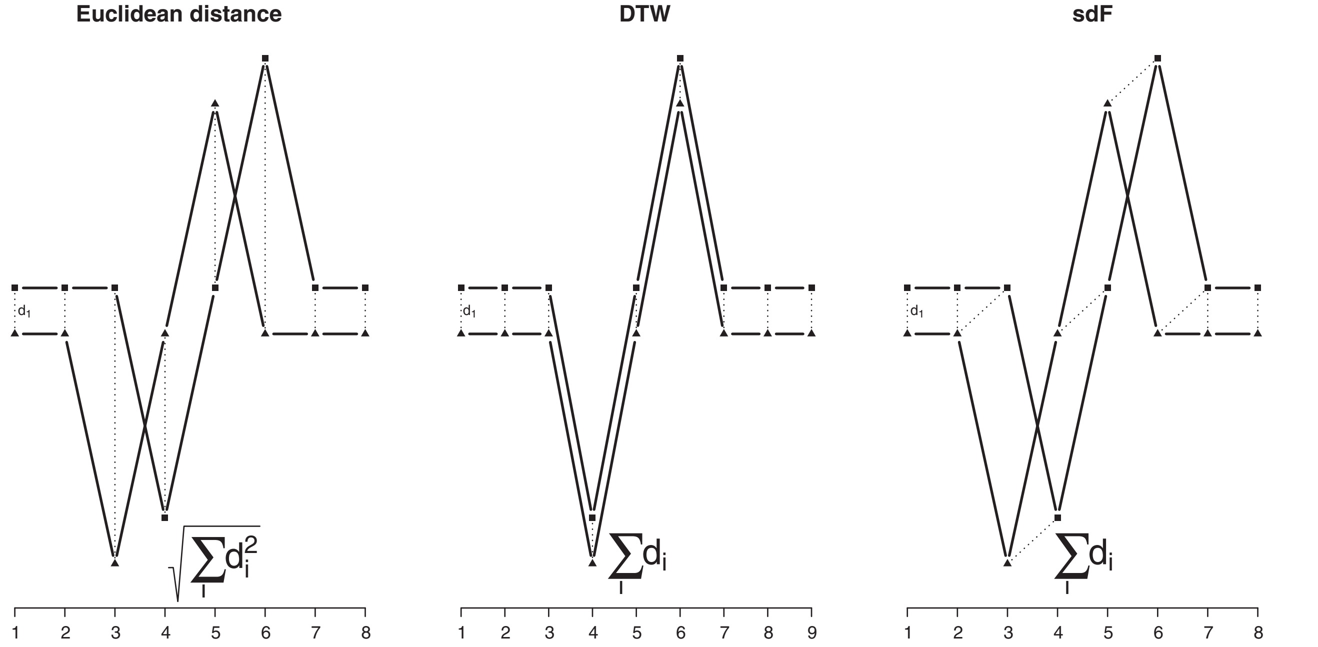


FIGURE 4 Schematic calculation principles for the 3 dissimilarity measures: Euclidean distance (left), dynamic time warping (middle), and generalized summed discrete Fréchet dissimilarity with $\lambda = 1$ (right). The Euclidean distance considers them pointwise. DTW first computes the optimal alignment before summing the pointwise differences on the y axis. sdF also computes an optimal alignment before summing the pointwise differences on the plane

measures are also "elastic" yet able to consider time shifts. As for DTW dissimilarity, they have a high calculation cost and also need a specific averaging process to provide cluster centroids. While the Euclidean distance requires two trajectories of the same length, an advantage of elastic dissimilarity such as DTW and sdF is that they can deal with trajectories having different numbers of points. Definitions 4 and 5 can be applied to time series of different lengths thanks to coupling, as recalled in Definition 2 (Appendix B.1).

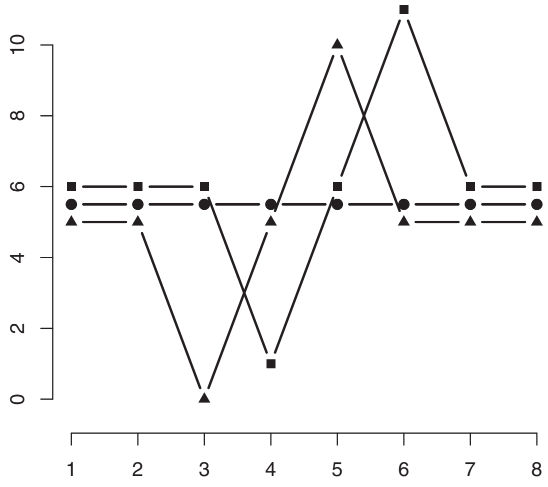
The parameter λ of the sdF dissimilarity can be understood as a time scale parameter. When $\lambda = 1$, the generalized discrete Fréchet distance becomes the classical discrete Fréchet distance and a difference of one unit on the variable of interest (the adhesion value here) is equivalent to a time lag of one time measure. A small value for λ gives more importance to differences in the variable of interest whereas a high value gives more importance to time lags (see Genolini et al[7] for more details). We note that the generalized summed Fréchet dissimilarity with $\lambda = 0$ matches the DTW measure. We have tested several values in a previous simulation experiment, and the results were not particularly sensitive to the choice of λ so we implemented a single value for this parameter. Within the context of our study, a difference of one hour between two trajectories was more clinically important than a time lag of one day, thus values inferior to 1 were preferred. We wanted to consider a constant difference of 4 hours between two trajectories as equivalent to a 15-day lag. Thus we set $\lambda = \frac{4}{15}$.

Figure 5 represents a third additional flat trajectory and the corresponding values for each dissimilarity and each pair of trajectories. The initial trajectories, marked with triangle and square are the farthest with the Euclidean distance, while they are the closest with DTW and sdF dissimilarities. This is explained because these trajectories share the same shape, with a time lag and only a difference of one unit on the y axis.

We were interested in how these dissimilarities manage the temporal shifts and find overall trends in the context of CPAP time series, which have high variability and many discontinuities due to the zeros. Especially, we asked if warping searches at any price to align the irregularities of the time series at the expense of grouping those with close trends.

3.2 | Linkage strategies

After the choice of a dissimilarity measure, the next step is to construct the hierarchy of the partition with a linkage strategy. This defines the dissimilarities between clusters from the individual pairwise dissimilarity matrix. The three strategies



Dissimilarity	Euclidean	DTW	sdF
d(▲, ●)	7.21	13.00	13.00
d(■, ●)	7.21	13.00	13.00
d(▲, ■)	10.39	9.00	11.07

FIGURE 5 Examples of three fictive trajectories and the corresponding dissimilarities

we considered were "average," "complete," and "Ward," among the most used and are described below. Definitions and formula are given in Appendix B.3.

The "average" linkage strategy defines the dissimilarity between two clusters as the average dissimilarity between each point of the first cluster and each point of the second cluster.

The "complete" linkage strategy defines the dissimilarity between two clusters as the largest dissimilarity between two subjects in the two clusters. Hence this strategy avoids merging two clusters if they contain time series strongly dissimilars.

The "Ward" method[11] consists of successively merging the two clusters which results in a minimal increase of the total within-cluster sum of squares. It is based on distance between objects and cluster centroids and was initially designed to work with the Euclidean distance and the arithmetic mean as centroid. This method has been extended to be convenient when considering other dissimilarities[12] and moreover, the Ward algorithm can be run using only the dissimilarity matrix[13] without use of centroids.

3.3 | Choosing the number of clusters

After constructing the partition hierarchy, the following step is to define the partition by choosing the number of clusters. An internal CVI computes a score from a partition of the dataset in terms of a dissimilarity measure. In practice, the indices are computed for the partitions obtained with different numbers of clusters and compared, allowing one to choose the "best" number of clusters. The CVIs differ by their definition of what is a good partition. Some of them work with a subjective decision like a knee point. We focused on objective CVIs, such that the higher or smaller the CVI is, the better the partition. Most indices are based on two criteria: compactness and separation. They have different properties,[14] but as far as we know there is no study examining them in the context of time series clustering. Here we considered the six following indices: Calinski-Harabasz (CH),[15] COP,[16] Davies-Bouldin (DB),[17] Modified Davies-Bouldin (DB*),[18] Dunn (D),[19] and Silhouette (Sil).[20] Note that CH, DB, DB*, and COP indices require to compute cluster centroids and CH also needs a global centroid.

3.4 | Centroids

The centroid of a set of objects is an object at the center of the set, related to a dissimilarity measure. Determining a centroid is a way of summarizing the set in terms of one object. Given a dissimilarity measure d, the centroid R_C of the set of objects C is defined by $R_C = \text{argmin}_{\xi \in \Xi} \sum_{x \in C} d(x, \xi)$, where Ξ has to be chosen. When d is the Euclidean distance, the centroid corresponds to the arithmetic mean at each time point of the time series. However, the classical mean is not the best centroid for every dissimilarity, especially with elastic measures. With these kinds of dissimilarity, due to the warping preceding the computation of the dissimilarity, the centroid may be of different length and time points from the

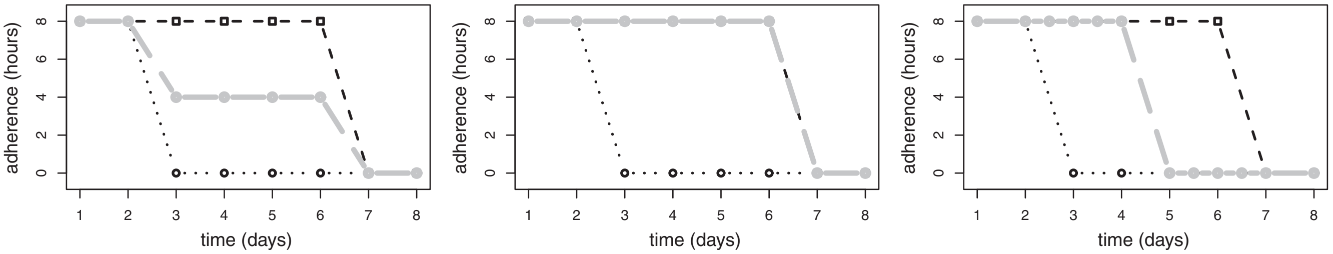


FIGURE 6 Examples (in gray) for the three centroids: The arithmetic mean (left), dynamic time warping barycenter averaging (DBA) (middle), and the Fréchet mean (right) on two fictive time series (dotted lines)

summarized time series. DTW barycenter averaging (DBA)[21] is a global averaging strategy providing centroids for groups of time series, well adapted to the DTW dissimilarity. The Fréchet mean[7] is an averaging strategy compatible with the sdF dissimilarity. These two strategies are heuristic and depend on an initialization step such that a distinction exists between the centroid algorithm and the centroid itself. Unlike the arithmetic mean, both the DBA and the Fréchet mean provide centroids with shapes similar to the individual time series, but at the price of a higher calculation cost. Figure 6 shows two fictive adherence time series over 8 days with application of the arithmetic mean, DBA, and Fréchet mean.

Another possibility is the medoid. The medoid of a group C is the centroid when $\Xi = C$. It is the object of C which minimizes the average dissimilarity to all other objects in C. The medoid is compatible with every dissimilarity measure.

3.5 | Software and packages

Within the R software environment (www.r-project.org), the package "dtwclust"[22] enables the clustering of time series with HAC among other methods. It includes the Euclidean distance and the DTW dissimilarity, with the DBA algorithm. It is possible to customize both the dissimilarity measure and the centroid process. The package "kmlShape"[7] includes the k-means algorithm with the sdF dissimilarity and the Fréchet mean. Combining these packages makes it possible to cluster time series with the sdF dissimilarity and using the HAC algorithm. "dtwclust" also includes the computation of some external CVIs such as the adjusted Rand index (ARI) and the 6 previously cited internal CVIs. We used this package for the simulation study.

4 | SIMULATION STUDY

In the simulation study, we compared the performance of the HAC algorithm using the three linkage strategies ("average," "complete," "Ward") and the three dissimilarity measures (Euclidean, DTW, sdF), that is, nine different clustering methods. The six internal CVIs mentioned in Section 3.3 were tested. R code is available upon request.

4.1 | Data generation process

Artificial data sets similar to real CPAP adherence time series were simulated following the results of Babbin et al,[23] in which clustering with HAC, the Euclidean and the Ward linkage strategy were applied to six month CPAP adherence series from 128 subjects. A partition with four clusters was obtained, called groups B, C, D, and E. As non-users and individuals stopping therapy early were excluded from their study, we added an extra group called "A" to our simulation study, corresponding to patients who dropped out, that is, completely stopped using CPAP. The length of the simulated time series was $T = 91$. Figure 7 shows the basic curves (see definition below) for each group. Simulated data were then obtained by adding white noise with variance σ^2.

Group A, dropout patients, was simulated with the model

$$X_{At} = \mathbb{1}_{t<\gamma} \times (5.5 + \epsilon_{At}) - \mathbb{1}_{\gamma-\tau\leq t<\gamma} \times \frac{5.5 \times (t - (\gamma - \tau - 1))}{\tau + 1},$$

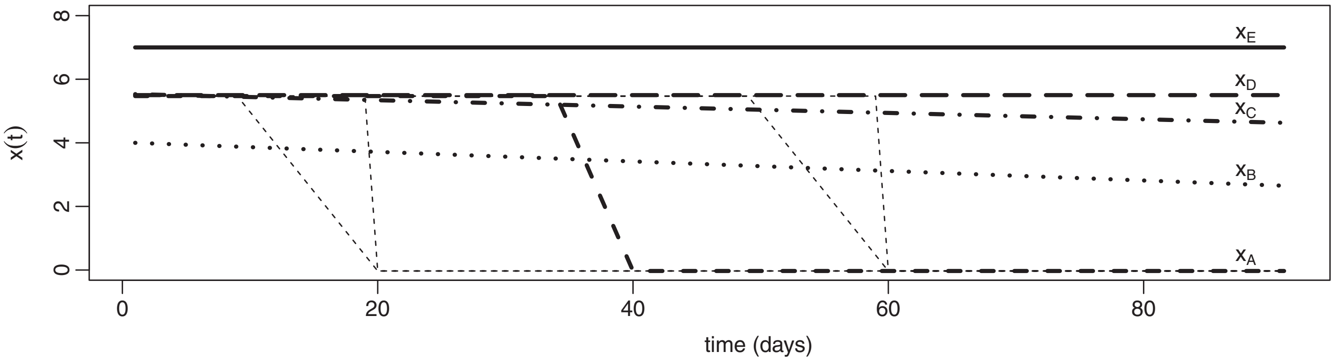


FIGURE 7 Non-noisy trends of groups A (dashed line), with 5 declensions varying the length of the phase of dwindling use and complete dropout, B (dotted), C (dots + dashes), D (long dashes), and E (solid) used for the simulation study

with an initial phase $[0, \gamma - \tau]$ with a constant trend, a phase of dwindling use $[\gamma - \tau, \gamma]$ and a "zero" or stopped phase $[\gamma, T]$, where γ and τ follow discrete uniform distributions $[20, 60]$ and $[0, 10]$ respectively. ϵ_{At} is white noise with variance σ^2. The four other groups were defined by:

$$X_{Bt} = 4 - 0.015t + \epsilon_{Bt},$$
$$X_{Ct} = 5.5 - 0.01t + \epsilon_{Ct},$$
$$X_{Dt} = 5.5 + \epsilon_{Dt},$$
$$X_{Et} = 7 + \epsilon_{Et},$$

where $\epsilon_{Bt},\ \epsilon_{Ct},\ \epsilon_{Dt},\ \epsilon_{Et}$ were white noise of variance σ^2.

To create data close to real-life data, four parameters were introduced: (1) group overlap, (2) presence of outliers, (3) variance of noise, and (4) null adherence values. Due to the proximity of groups C and D, including them or not in the datasets was a way to create group overlap. When the 5 groups were present in a sample (overlap), we generated 150 time series per group. For samples constituted only of groups A, B, and E (no overlap), each group contained 250 subjects. The second parameter was the presence or not of outliers. When included, they added 75 (10%) supplementary curves, each characterized by a first value and a slope. The model for an outlier trajectory was $X_t = \alpha + \beta t + \epsilon_t$, where α and β follow uniform distributions of $[2, 8]$ and $[-0.05, 0.05]$ respectively, and ϵ_t was white noise of variance σ^2. The variance of the noise σ^2 was either 2.5 or 6.5. These values corresponded approximately to the first and third quartiles of the individual variances estimated on real datasets (see Figure A1). The last parameter aimed at generating null adherence values. A null value was introduced when the simulated adherence value was below the censorship limit ρ, either fixed at 0 or at 1.5. In the real dataset, 4.1% of the 77 168 values were non-null and below 1.5 hours. Examples of simulated adherence trajectories from each group are shown in Figure 8.

These choices led to 16 parameter combinations. A hundred samples of 750 or 825 time series were simulated for each of them. Each of these datasets was clustered using the HAC algorithm, with either the "average," "complete," or "Ward" linkage strategies, and either the Euclidean distance, DTW dissimilarity or the generalized summed discrete Fréchet dissimilarity. The performances of these nine different clustering methods were compared using the methodology described below.

4.2 | Comparison of clustering methods

For a sample j, the true partition is p_j. In the case of a sample with outliers, each extra-group time series is considered as its own group. To evaluate the quality of the partition provided by a clustering algorithm when the true partition is known, an external CVI can be used returning a number in the range $[0, 1]$ (or $[-1, 1]$) such that the greater the similarity among partitions the nearer the value is to 1. Several external CVIs exist and we chose the ARI.[24] Let $h_{j,m}$ be the dendrogram obtained on sample j with the clustering method m, and $p^k_{j,m}$ the corresponding extracted partition with k clusters. A naive evaluation considers only the partition with the true known number of groups, and compares it with the real partition.

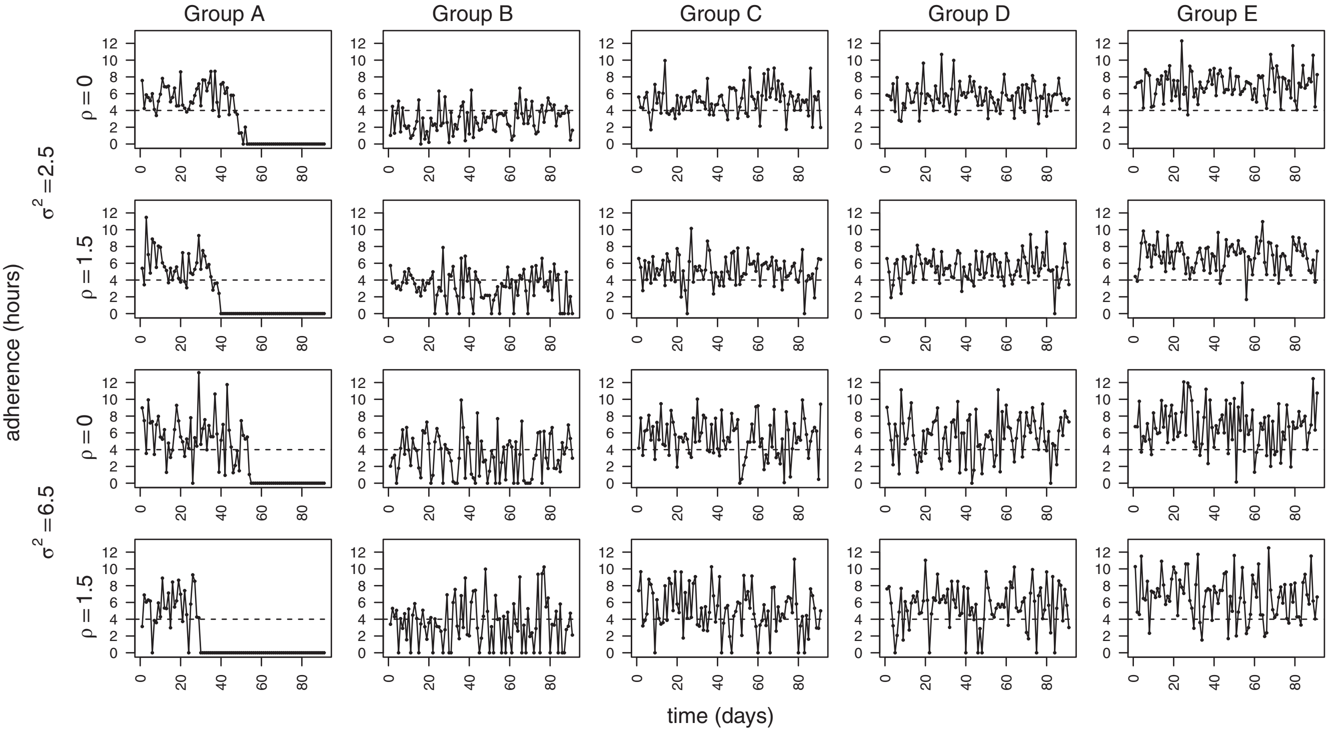


FIGURE 8 Examples of simulated trajectories. Each column stands for a group A, B, C, D, and E from left to right. Each combination of noise variance (σ^2) and censorship limit (ρ) are shown in rows

This approach assumes that among partitions resulting from a clustering method, the best one has the true number of clusters.[8] However this assumption does not hold, Figures C1 and C3 (in Appendix C) show an example where the best partitions do not have the true number of clusters. Instead of looking at only the partition with the true number of clusters, we computed the set of the best number of clusters with respect to ARI, with k ranging from 2 to 20. We note $\widehat{K_{j,m}} = \text{argmax}_{k \in [\![2;20]\!]}\ ARI(p^k_{j,m}, p_j)$ this set. However $\widehat{K_{j,m}}$ can be either unique (only one partition of the dendrogram maximizes the ARI) or a set number of clusters achieving the maximal ARI value. Let us denote $\widehat{P_{j,m}} = \{p^k_{j,m} \mid k \in \widehat{K_{j,m}}\}$ the set with the best partitions for the dendrogram $h_{j,m}$. The performance of the method m on sample j is evaluated through $s_{j,m} = ARI(p, p_j)$ where p is one element of $\widehat{P_{j,m}}$. The classification is perfect when its best partition exactly matches the original one (ie, $s_{j,m} = 1$). To compare methods we computed their perfect classification rates (PCR). We also looked at the $s_{j,m}$ distributions through their mean ($\overline{s_{\cdot,m}}$) and standard deviation ($s_{s_{\cdot,m}}$).

After the choice of the clustering method, the next step was to determine the best number of clusters.

4.3 | Comparison of internal CVIs

Let us consider a dendrogram $h_{j,m}$ with its associated partitions $\{p^k_{j,m} \mid 2 \leq k \leq 20\}$, an internal CVI v and a centroid algorithm $\mathcal{R}$. We note d the dissimilarity measure that is related to the clustering method m.

Let $\widetilde{K^{v,\mathcal{R}}_{j,m}}$ be the set of value $k \in [\![2;20]\!]$ which optimizes $v(p^k_{j,m}, d, \mathcal{R})$ and $\widetilde{P^{v,\mathcal{R}}_{j,m}} = \{p^k_{j,m} \mid k \in \widetilde{K^{v,\mathcal{R}}_{j,m}}\}$. The performance of v with the centroid algorithm $\mathcal{R}$ and the method m on sample j is evaluated through $s^{v,\mathcal{R}}_{j,m} = \frac{\sum_{k \in \widetilde{K^{v,\mathcal{R}}_{j,m}}} ARI(p^k_{j,m}, p_j)}{|\widetilde{K^{v,\mathcal{R}}_{j,m}}|}$, the mean ARI between the true partition p_j and each element of $\widetilde{P^{v,\mathcal{R}}_{j,m}}$. The CVI v was considered to be a success on the sample j with clustering method m and centroid algorithm $\mathcal{R}$ when there was a partition maximizing ARI among the partition optimizing v (ie, $\widehat{P^m_j} \cap \widetilde{P^{v,\mathcal{R}}_{j,m}} \neq \emptyset$). We compared the performance of the internal CVIs with their success rates (SR) and also by comparing their $s^{v,\mathcal{R}}_{j,m}$ scores through their mean ($\overline{s^{v,\mathcal{R}}_{\cdot,m}}$) and standard deviation ($s_{s^{v,\mathcal{R}}_{\cdot,m}}$).

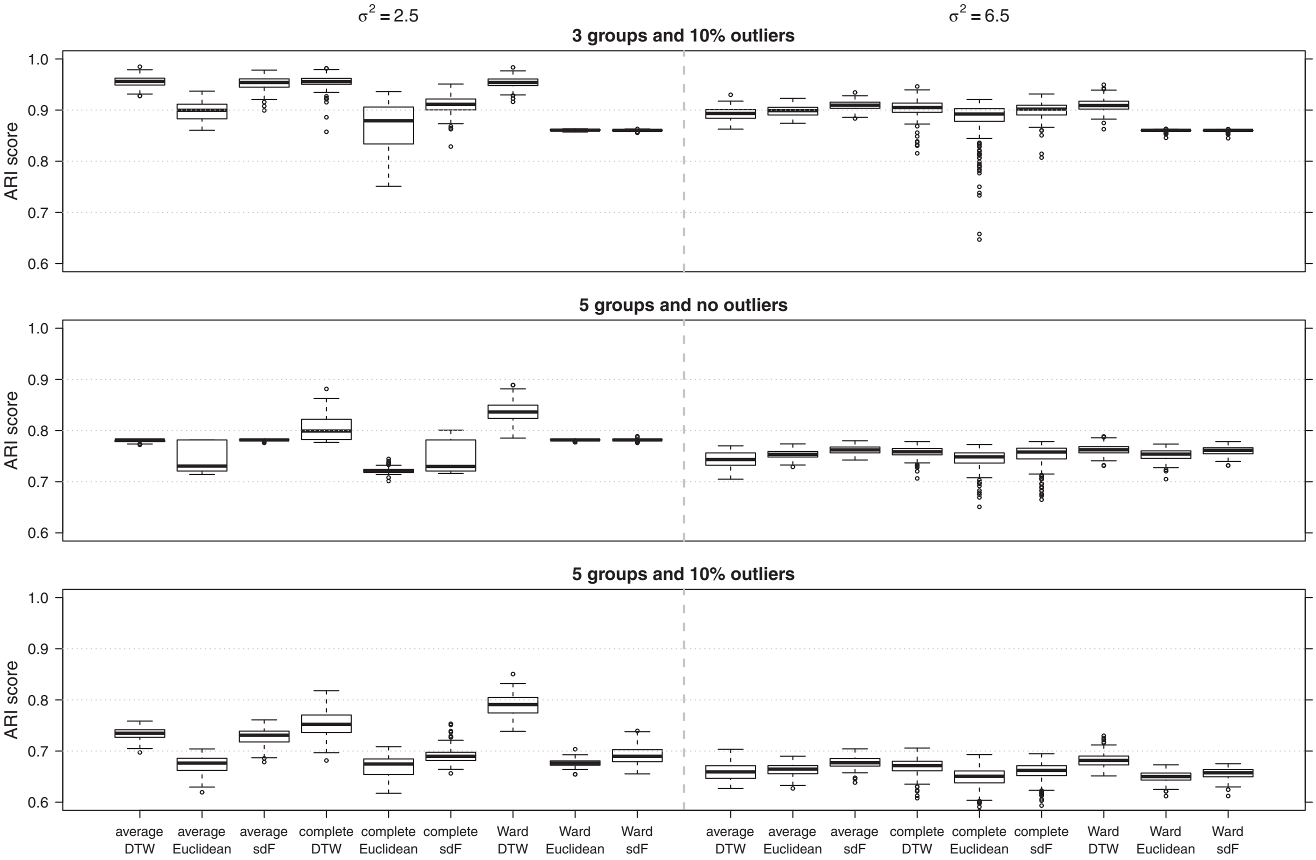


FIGURE 9 Simulation study: Boxplots of $s_{j,m}$ scores for each method in cases with 3 groups and 10% of outliers, with 5 groups without outliers, and with 5 groups and 10% of outliers. Boxplots are split according to the variance σ^2 making 6 different cases of 200 samples

4.4 | Simulation results

4.4.1 | Comparison of clustering methods

The simplest case is 3-group clustering. When there is no outlier, almost all methods reach near 100% perfect classification, except two of them. Average linkage with DTW dissimilarity and complete linkage with Euclidean distance exhibited rates of perfect classification of around 50% (see Table 1). With 10% of outliers, there was no perfect classification. Average and complete linkage strategies could produce a lot of clusters in the final partition (see Figure C2 in Appendix C), which, in practice, is a drawback. The distributions of $s_{j,m}$ scores with 3 groups and outliers, or with 5 groups without and with 10% of outliers are presented in Figure 9. With 5 groups and no outlier the Ward linkage strategy gave the best results in term of $\overline{s_{.,m}}$ (Table 1). In the case of low variance ($\sigma^2 = 2.5$), 5 theoretical groups without outliers (see middle plot in Figure 9), the DTW dissimilarity measure with the Ward linkage strategy clearly yielded the best scores even for 5 groups and 10% of outliers, whatever the variance (see bottom plot in Figure 9). In this configuration the DTW dissimilarity with the Ward linkage strategy gave the best mean $\overline{s_{.,m}}$ (see Table 1). We also noticed that there was a large range of values for the number of clusters in the final partition with average and complete linkage (see Figure C4 in Appendix C). The presence of outliers considerably influenced the mean $\overline{s_{.,m}}$ and the number of clusters in the resulting partitions (see Figures C1-C4 in Appendix C). A large variance σ^2 shrunk the global performance of the clustering methods and reduced the differences between methods. The censorship limit and thus a greater presence of null values had no impact on the clustering performances. This is important regarding its practical application to CPAP time series.

Finally, from the simulation study, the DTW dissimilarity used with the Ward linkage strategy provided the best results both in terms of $s_{j,m}$ score distributions and perfect classification rates. Each clustering method was applied to the same 1600 datasets. We ran 8 Wilcoxon tests comparing each combination to DTW-Ward. The highest P-value after Bonferroni correction was close to 4.99×10^{-73} showing that DTW-Ward was thus significantly better than the other combinations.

TABLE 1 Simulation study: Comparison of the methods according to group overlap (no overlap with 3 groups, overlap with 5 groups) and presence of outliers (400 samples per combination)

		3 groups					5 groups			
		No outliers			10% of outliers		No outliers		10% of outliers	
Linkage	Dissimilarity	PCR	$\overline{s_{\cdot,m}}$	$s_{s_{\cdot,m}}$	$\overline{s_{\cdot,m}}$	$s_{s_{\cdot,m}}$	$\overline{s_{\cdot,m}}$	$s_{s_{\cdot,m}}$	$\overline{s_{\cdot,m}}$	$s_{s_{\cdot,m}}$
Average	DTW	0.54	0.99	0.01	0.92	0.05	0.76	0.02	0.69	0.05
	Euclidean	0.97	1.00	0.00	0.90	0.02	0.75	0.02	0.67	0.01
	sdF	1.00	1.00	0.00	0.93	0.02	0.77	0.01	0.70	0.03
Complete	DTW	1.00	1.00	0.00	0.93	0.03	0.78	0.03	0.71	0.05
	Euclidean	0.53	0.94	0.06	0.88	0.04	0.73	0.02	0.66	0.02
	sdF	1.00	1.00	0.00	0.90	0.02	0.75	0.02	0.68	0.02
Ward	DTW	1.00	1.00	0.00	0.93	0.03	0.80	0.04	0.74	0.06
	Euclidean	1.00	1.00	0.00	0.86	0.00	0.77	0.02	0.66	0.02
	sdF	1.00	1.00	0.00	0.86	0.00	0.77	0.01	0.67	0.02

Note: Perfect classifications rates (PCR), mean $s_{j,m}$ scores ($\overline{s_{\cdot,m}}$), and standard deviation ($s_{s_{\cdot,m}}$).

TABLE 2 Simulation study: Success rates (SR), mean $s^{v,\mathcal{R}}_{j,m}$ scores ($\overline{s^{v,\mathcal{R}}_{\cdot,m}}$), and standard deviation ($s_{s^{v,\mathcal{R}}_{\cdot,m}}$) for each internal cluster validity index (CVI) depending on the number of groups and presence of outliers (400 samples for each combination)

	3 groups						5 groups					
	No outliers			10% of outliers			No outliers			10% of outliers		
CVI (centroid)	SR	$\overline{s^{v,\mathcal{R}}_{\cdot,m}}$	$s_{s^{v,\mathcal{R}}_{\cdot,m}}$	SR	$\overline{s^{v,\mathcal{R}}_{\cdot,m}}$	$s_{s^{v,\mathcal{R}}_{\cdot,m}}$	SR	$\overline{s^{v,\mathcal{R}}_{\cdot,m}}$	$s_{s^{v,\mathcal{R}}_{\cdot,m}}$	SR	$\overline{s^{v,\mathcal{R}}_{\cdot,m}}$	$s_{s^{v,\mathcal{R}}_{\cdot,m}}$
Silhouette	0.00	0.57	0.00	0.00	0.48	0.01	0.00	0.37	0.00	0.00	0.30	0.02
Calinski Harabasz (Medoid)	0.00	0.57	0.00	0.00	0.48	0.01	0.00	0.37	0.03	0.00	0.30	0.04
Calinski Harabasz (DBA)	0.01	0.57	0.04	0.00	0.48	0.01	0.00	0.37	0.01	0.00	0.30	0.02
Davies Bouldin (Medoid)	0.00	0.57	0.00	0.00	0.48	0.04	0.01	0.38	0.05	0.00	0.31	0.05
Davies Bouldin (DBA)	0.32	0.71	0.20	0.00	0.60	0.18	0.00	0.42	0.05	0.00	0.36	0.08
Modified Davies Bouldin (Medoid)	0.00	0.57	0.00	0.00	0.48	0.02	0.02	0.38	0.06	0.00	0.31	0.05
Modified Davies Bouldin (DBA)	0.11	0.62	0.13	0.00	0.52	0.12	0.00	0.42	0.05	0.00	0.36	0.08
Dunn	0.86	0.94	0.15	0.76	0.86	0.16	0.28	0.75	0.06	0.34	0.67	0.09
COP (Medoid)	0.00	0.32	0.02	0.00	0.39	0.03	0.00	0.38	0.02	0.00	0.43	0.05
COP (DBA)	0.00	0.33	0.02	0.00	0.40	0.03	0.00	0.38	0.02	0.00	0.43	0.05

Note: CVIs for which a centroid is needed are displayed on two lines. The first line corresponds to the CVI computed with the medoid and the second corresponds to the CVI computed with dynamic time warping barycenter averaging.

4.4.2 | Comparison of internal CVIs

The internal CVIs were evaluated with the clustering based on Ward linkage and DTW dissimilarity (see Table 2). Except for the Dunn index, and whatever the simulation parameters, all internal CVIs performed very poorly in identifying the number of clusters which maximizes the ARI index. The Dunn index was the most effective with high success rates with 3 groups but somewhat lower with 5 groups. This was confirmed by the mean $\overline{s^{v,\mathcal{R}}_{j,m}}$ scores (Table 2). Based on this simulation study with the clustering method previously selected, the best internal CVI to choose the number of clusters for CPAP adherence series is the Dunn index.

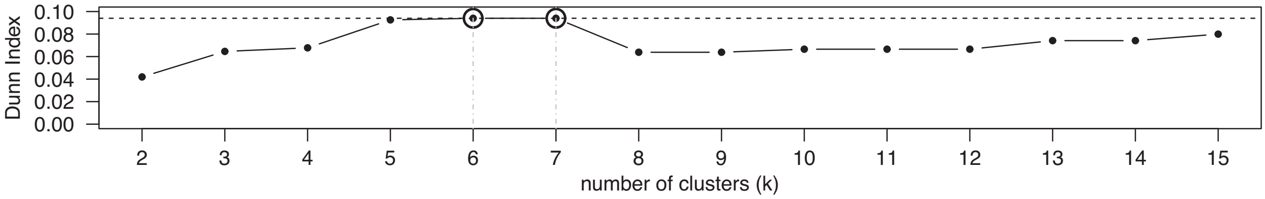


FIGURE 10 Real CPAP data: Dunn index values by number of clusters

5 | REAL-LIFE CPAP ADHERENCE CLUSTERING

Our aim was to explore individual CPAP adherence data using cluster analysis, to reveal typical patterns. According to the simulation results, the data presented in Section 2 were clustered with the Ward linkage strategy and the DTW dissimilarity, and the number of clusters was selected with the Dunn index.

Partitioning giving from 2 to 15 clusters were considered. Figure 10 represents the Dunn index values with the number of clusters. Partitions with 6 or 7 clusters both maximized the internal CVI. For sparsity reasons, the partition with the lowest number of clusters was chosen first. Figure 11 shows the median adherence trajectories for the 6 clusters. The first cluster contains 105 subjects with a median use of 4 hours per night at the beginning of therapy; gradually increasing nearly 8 hours per night at the end of 3 months. The variances in the time series are high, especially at the beginning of treatment, and patients have some zeros. The second cluster comprises 151 highly adherent users with a median trajectory stabilizing at around 8 hours use per night. Zero values are rare and the individual variances are low. The third cluster groups 29 patients who stopped CPAP. Individual variances are high and many zeros appear, especially towards the complete discontinuation of treatment. The fourth cluster contains 150 subjects with good adherence, a median trajectory of nearly 6 hours per night, individual variances are quite low and few zeros appear. Cluster 5 brings together 178 non-adherent users, a median trajectory beginning at 2 hours that decreases. Individual variances are low to moderate and several individual trajectories contain many zeros. The sixth and largest cluster contains 235 subjects, the median trajectory is about 4 hours, but individual variances are moderate and some zeros appear. Clusters 2, 4, and 6 represent subjects with predominately stable patterns but at different levels. One might consider that the difference between these clusters is due to sleep duration. However, for clusters 2 and 4, the difference in the median mean adherence is almost an hour and a half. This corresponds to a complete sleep cycle, which has clinical meaning. Cluster 6 presents smaller proportions of nights with usage at over 4 hours, with higher rates of zero use and greater variances, indicating that cluster 6 also differs from clusters 2 or 4 in usage regularity.

Clustering was performed using a HAC method. Thus the partition among 6 clusters is obtained from the partition into 7 clusters and then merging two clusters. The cluster 2 results from merging 2.1 and 2.2, containing respectively 40 and 111 subjects. Figure 12 shows their median trajectories. Really good users are thus spread between two clusters where median trajectories remain stable near 8 hours per night for both clusters. The difference between these two clusters is in the regularity of adherence. Individual medians are higher in cluster 2.1 than in cluster 2.2, but with higher variances and more occurrences of zero. Keeping in mind that the main motivation of this study was personalized support for subjects at risk of dropping out or being insufficiently adherent, this distinction is not relevant. The 6-cluster partition was finally retained. Table 3 also describes these clusters and compares them. The independence between these clusters and continuous variables was tested using the Kruskal-Wallis test, and the independence with the discrete variables was tested using the chi-square test. The significance level of the tests is fixed at 5%. We note that belonging to one cluster or another is not independent of the age of the patient, nor of other individual adherence characteristics presented on this table (mean, standard deviation, rate of zeros and proportion of nights with adherence over 4 hours).

The median trajectories are a way to represent the 6 clusters but do not properly show individual characteristics in terms of null values and dropouts for example. This is especially so for clusters 3 and 5 that group subjects with many zeros, the first because of dropouts and the second because of poor and irregular adherence. Boxplots in Figure 13 give more information, showing for each cluster the distribution of the day with the first zero and of the day of completely stopping CPAP, that is, dropout, when applicable. The distributions of these two characteristics for the overall population are also shown in the histograms in Appendix A (Figure A1). The distribution taking into account the day of the first zero showed that subjects in the "dropout" cluster (cluster 3) have zero use at later times than subjects in cluster 5 (poor adherers). The distribution that considers the time of dropout revealed two facts about the resulting partition. The first was that

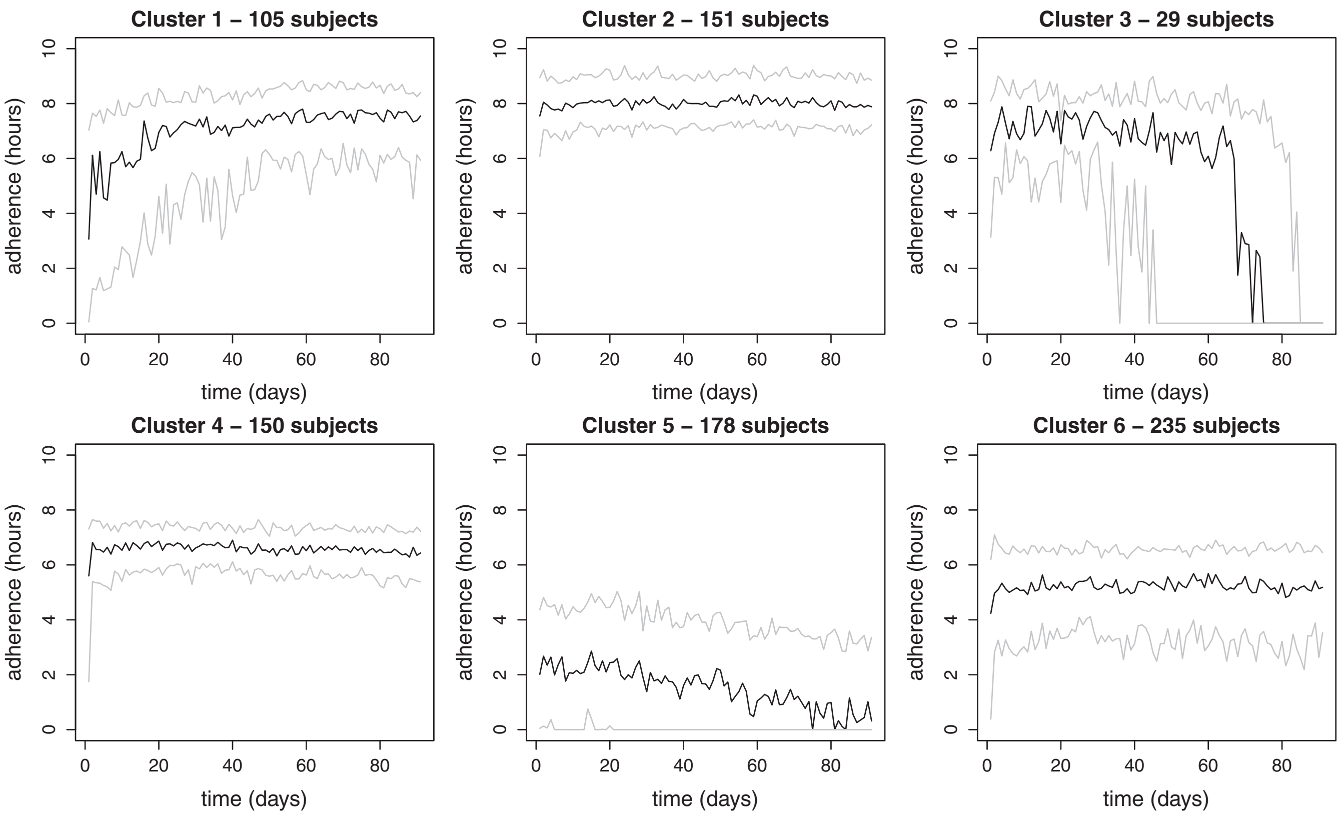


FIGURE 11 Real CPAP data: Representatives of the 6 clusters with median trajectories (black) and quartile trajectories (gray). Median and quartile trajectories are calculated pointwise

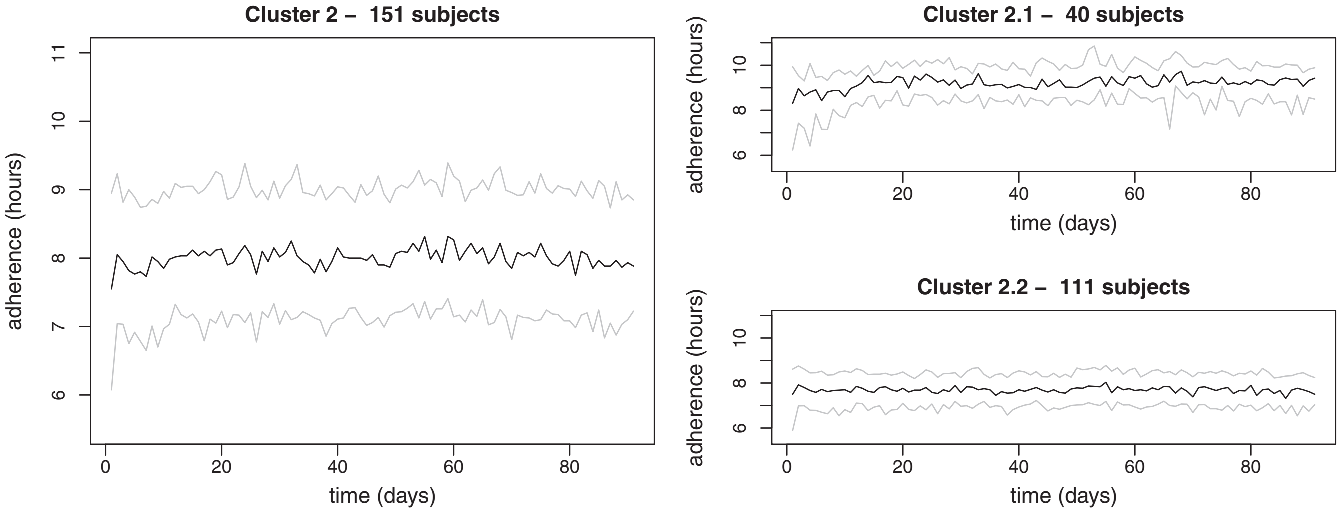


FIGURE 12 Real CPAP data: Median (black) and quartile (gray) trajectories for cluster 2 (left) from the partition with 6 clusters and for clusters 2.1 and 2.2 (right) from the partition with 7 clusters. Median and quartile trajectories are calculated pointwise

TABLE 3 Real CPAP data: Descriptive statistics for the selected population and comparison of the 6 clusters

	Male sex	Age (year)	Body mass index (kg/m^2)	Apnea hypopnea index (event/hour)	Mean adherence (hour)	Standard deviation of adherence (hour)	Rate of zeros (%)	Proportion of adherence >4 hours (%)
Whole population								
N = 848	533 (62.9%)	59 [50; 69]	30.4 [26.7; 34.6]	36.0 [30.0; 50.1]	5.5 [3.8; 6.7]	2.1 [1.5; 2.6]	6.6 [1.1; 19.8]	78.6 [51.6; 93.4]
Missing values	0 (0%)	2 (0.2%)	66 (7.8 %)	204 (24.1 %)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Cluster comparison, n (%)								
Cluster 1, 105 (12.4%)	68 (64.8%)	62 [45; 70]	28.0 [25.6; 33.1]	37.5 [30.0; 54.5]	6.3 [5.5; 7.0]	2.9 [2.4; 3.3]	8.8 [4.4; 14.3]	81.3 [68.1; 87.9]
Cluster 2, 151 (17.8%)	91 (60.3%)	60 [52; 70]	30.8 [27.1; 35.2]	39.0 [31.0; 58.0]	7.8 [7.3; 8.3]	1.4 [1.0; 1.8]	0.0 [0.0; 1.6]	97.8 [95.6; 98.9]
Cluster 3, 29 (3.4%)	15 (51.7%)	58 [45; 69]	30.9 [27.9; 36.0]	35.0 [30.0; 40.0]	5.5 [3.0; 6.4]	2.9 [2.6; 3.4]	22.0 [9.9; 53.8]	72.5 [40.7; 85.7]
Cluster 4, 150 (17.7%)	90 (60.0%)	63 [54; 72]	30.0 [27.0; 33.1]	35.0 [30.0; 48.0]	6.4 [6.0; 6.6]	1.7 [1.4; 2.0]	1.1 [0.0; 3.3]	92.3 [88.2; 95.6]
Cluster 5, 178 (21.0%)	120 (67.4%)	57 [48; 65]	30.9 [26.8; 36.4]	34.5 [29.0; 45.6]	2.2 [1.3; 3.2]	2.0 [1.5; 2.5]	33.5 [17.6; 53.8]	23.6 [8.8; 39.3]
Cluster 6, 235 (27.7%)	149 (63.4%)	58 [50; 67]	30.2 [26.2; 34.7]	36.5 [30.0; 51.0]	4.8 [4.2; 5.3]	2.3 [2.0; 2.8]	8.8 [3.3; 18.7]	70.3 [58.8; 79.1]
P-value	0.510	**0.003**	0.091	0.133	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Note: Continuous variables are summarized using median [interquartile range] while discrete variables are summarized using frequency (proportion). The independence between these variables and clusters was tested using the Kruskal-Wallis test (continuous variable) and the chi-square test (discrete variable). The p-values under the significance level fixed at 5% are given in bold font.

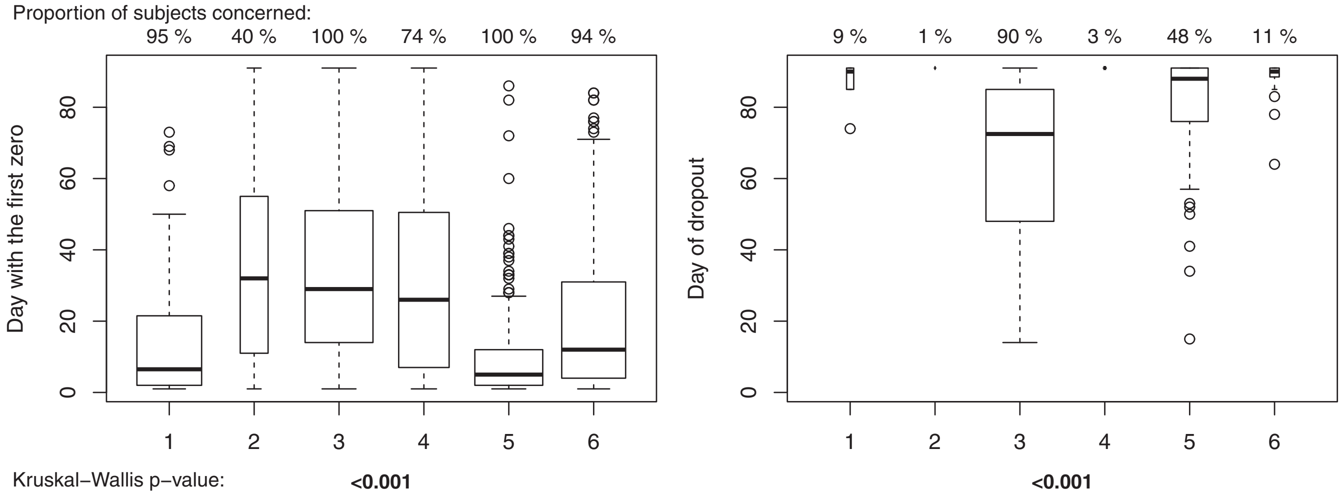


FIGURE 13 Real CPAP data: For each cluster, distribution of the day of the first zero (left plot) and distribution of the day of dropout (right plot). The proportions of subjects concerned are given above each boxplot and the widths of the box are proportional to these proportions. Clusters are compared using the Kruskal-Wallis test

more subjects in cluster 3 (90%) actually end their trajectory with a zero, than in cluster 5 (48%). The second was that the time when complete dropout occurred is earlier in the "dropout" cluster. Some patients who eventually stop CPAP, who begin with poor adherence are thus grouped in cluster 5. This is not visible using the median and quartile trajectories because these patients are grouped with subjects still continuing their CPAP therapy at the end of the third month. This highlights the difficulty in giving representatives of clusters based on time series clustering alone and advocates the need to use several tools. Figure D1 in Appendix D shows these clusters with their medoids.

Nevertheless, the comparison of the clusters gave significant differences for characteristics involving zero values (rate of zero [see Table 3], day of the first zero [Figure 13], and time of complete dropout [Figure 13]). This observation underlines that the distribution of zeros have been taken into account by the clustering algorithm, which gives reassurance in the resulting partition.

The two clustering approaches which obtained the nearest performances to that of the DTW-Ward combination in the simulation study were sdF-average and DTW-complete combinations. They were applied to the real-life dataset fixing the number of clusters at six for comparability reasons. Both approaches produced partitions with clusters of unbalanced sizes. SdF-average clustering identified four clusters with very few subjects (from one individual to seven) and DTW-complete produced one large cluster (385 subjects) and two small clusters (30 and 21 subjects). Clustering with the DTW measure and Ward linkage tended to identify more equilibrated clusters. Hence the resulting clusters are less susceptible to be dependent on a given dataset.

6 | DISCUSSION

In this work, we investigated complete clustering procedures primarily intended for use with CPAP adherence time series in an HAC framework. We considered the choice of dissimilarity measure taking into account several linkage strategies comparing an original dissimilarity measure, the generalized summed discrete Fréchet dissimilarity, with the classical Euclidean distance and the DTW dissimilarity that is widely used in time series clustering. We also examined the question of the selection of the number of clusters, using the six following internal clustering validation indices: Calinski-Harabasz, COP, Davies-Bouldin, Modified Davies-Bouldin, Dunn, and Silhouette.

In a simulation study we have shown that a combination of DTW dissimilarity, Ward linkage and the Dunn index provided the best results within the HCA context. Fictive datasets with parameters that included the presence of extra-group outliers, the variance of noise and the number of zeros were used. We found the presence of outliers to be the most influential parameter, reducing the performances of the clustering algorithm (using the Ward linkage and the DTW dissimilarity) and that of the Dunn index with this algorithm. Next, the variance of the noise decreased the performances of the internal CVI, with and without the presence of outliers and whatever the number of zeros, and also that of the clustering algorithm, except in the case of 3 groups without outliers where performance was not affected. An increase in zeros did not change the performance of the clustering algorithm but substantially diminished the CVI. Nevertheless, the effects of these parameters were minimal so the whole clustering method appears stable enough to give confidence in the partition we would obtain with real data.

Some improvements in the design of the simulation could be made regarding the clinical question and the specificity of data to cluster. A first is to consider the timing of the phase of diminishing CPAP use and of complete dropout in group A as simulation parameters and to measure their impact on clustering performances. This question is related to the time scale parameter λ of the summed discrete Fréchet dissimilarity. Only a single value of λ was used, however the performances of the clustering methods based on sdF dissimilarity could vary with the supplementary simulation parameter previously mentioned and with other values of λ. We simulated only one group with a shape recognizable by elastic dissimilarity measures. An additional group with a zero phase, like subject 8 in Figure 3, could be simulated, also varying the number of days with such phases.

From a methodological point of view, it would be interesting to use partitional clustering with the sdF dissimilarity and Fréchet mean centroids, but the computational cost is too high. A further point concerns the choice of the number of clusters. Except for the Dunn index, all internal CVIs we implemented in the simulation study gave very poor performances in selecting partitions close to the theoretical ones. This was even true with the Euclidean distance or sdF dissimilarity. It appears that the CVIs we tested are not suitable for time series clustering. More research is needed to develop new internal CVIs that are appropriate for use with time series.

The application of HAC with Ward linkage, DTW measure and the Dunn index provided six clusters, ranging from a small cluster (N = 29 subjects) in which patients stopped the therapy early on, to a cluster (N = 105) in which patients increased their adherence over time. Other clusters included extremely good users (N = 151), good users (N = 150), moderate users (N = 235), and poor adherers (N = 178). Regarding the previous study on CPAP data conducted by Babbin et al[23] we found two supplementary and considerably different CPAP clusters, both characterized by their shape. As expected, one was users who discontinued the therapy and the other contained subjects who began with only moderate use and greatly increased CPAP use by the end of the 3 months. This was probably due to two main reasons. First because we considered non-users and zero values. It is obviously necessary to define clusters having extended periods of non-use. The second reason may be the dissimilarity we used that aligns time series to find similar shapes over time. Application of the Euclidean distance and Ward linkage strategy, as used by Babbin et al[23] along with the Dunn index, to our real dataset gave 8 clusters, but did not produce a "dropout" cluster. The subjects who dropped out were dispatched into several clusters depending on the date of dropout, but the date is not of clinical importance. More importantly, the six clusters identified using HAC with Ward linkage, DTW measure and the Dunn index made clinical sense and allow both clinicians and home care-providers to ensure a better allocation of resources by individualizing patient management during the stabilization phase.

These results were obtained considering a 3-month time window after treatment initiation for the CPAP trajectories. Applying the same clustering methodology to only the first 2 months of follow-up would not have recognized the dropouts. This underlines how the choice of time window impacts on the resulting clusters. An interesting research question would be to apply the clustering to the same trajectories varying the time window so as to identify the moment after which the clusters become quite similar over time, and giving an estimate of the time required before CPAP adherence stabilizes.

The description of CPAP adherence trajectories from individual trajectories has previously been attempted.[23] However, to our knowledge the work we report here is the first investigating clustering methods that consider the specificities of CPAP data, that is, high variability and many discontinuities, which render them different from classical time series.

In this study, we highlighted some advantages of using the DTW dissimilarity for clustering trajectories: (1) the possibility to apply it to trajectories of different lengths; (2) its ability for recognize shapes, illustrated by grouping together dropouts whatever the dates of dropout were. However, this was at the expense of an increase in computation time.

We also investigated the data emission process for the different models of CPAP device, so as to formulate hypothesis enabling us to distinguish zero use from missing values. One hypothesis was that for the devices U and V there were no missing values. This seems reasonable, but this topic merits complementary work to confirm the reliability of the data supplied by each CPAP model; especially null and missing values. A special focus on missing values could be a future research question, avoiding the patient exclusion seen in this study. It could use an imputation strategy or a dissimilarity measure that takes otherwise excluded patients into account. Another direction of study would be to look at the zero values because they could furnish information on individual adherence behaviors. First, an extension of this work would be to search for a smoothing strategy that preserves the zero values, without replacing them with positive averaged values; and a complementary approach would be to model the occurrence of zero values to better define poor adherence profiles.

A limitation to these results is that some clusters, such as extremely good users and good users, may differ due to their sleep duration rather than to adherence behaviors. Currently, sleep duration is mainly subjectively reported at treatment initiation and, to our knowledge, there is no dataset, which combines both daily sleep duration and daily CPAP adherence. For this study, we used the criteria of the French National Health Insurance System for CPAP reimbursement, which is based only on the number of hours of CPAP use. However, the method used here for hours of CPAP adherence could be used in the same way for the daily ratio of CPAP use to sleep duration. This type of data is currently unavailable, but the development of connected health applications might make such more complete data available in the future.

Beyond application to CPAP adherence data, the clustering algorithm combining the DTW dissimilarity and HAC with the Ward linkage strategy, could be extended to all time series containing discontinuities and high variability which are common in medical area. These methods can be applied in sleep apnea for other data generated from CPAP telemonitoring: the residual number of apnea events under CPAP (residual apnea hypopnea index) or air leaks during CPAP use. These data patterns are also observed during the initiation period of non-invasive ventilation in patients with chronic obstructive pulmonary disease. In other fields, daily data recorded by activity trackers, accelerometers, pedometers, smart-phone applications, or wearable medical devices can present characteristics of variability and discontinuities. The necessity for unsupervised clustering was highlighted in previous studies.[25,26] Another field of application of such methods is the monitoring of blood markers, if collected at regular time periods, such as the residual concentration of immunosuppressant drugs after an organ transplant. Indeed, for data presenting the same specific problematics, clustering them through the DTW dissimilarity and HAC with the Ward linkage strategy might help to identify trajectory patterns at the beginning of a treatment or an intervention.

To conclude, our initial clinical objective was to enhance CPAP adherence follow-up for patients newly diagnosed with OSA. The DTW dissimilarity and HAC with the Ward linkage strategy seems well adapted to this approach in the field of OSA patient management. The resulting clusters showed clinical relevance and provided a richer description of the adherence behaviors than mean adherence values do. Comparing CPAP adherence clusters could confirm the effectiveness of a therapy and help to define more precisely efficiency thresholds.[27] The application of such methods is in line with the development of “P4” medicine (personalized, predictive, preventative, and participatory) in the field of OSA identifying specific trajectories of patient treatment..[28] Moreover, this clustering method can be applied to a broad spectrum of regularly collected data in different medical fields so as to improve personalized medicine.

The next step will be to use individual data collected at CPAP prescription to predict whether a given patient belongs to one cluster or another and to characterize these clusters. Such a predictive approach would help doctors and home care providers to lighten the follow-up of patients who are predicted to be adherent and would allow them to concentrate their efforts on subjects at risk of dropping out or being poor adherers. Another clinical application of this work would be to evaluate the impact of belonging to a particular cluster on the improvement in daytime symptoms and on the risk of comorbidities.

ACKNOWLEDGEMENTS

GBB, SB, ALS, and JLP are supported by the French National Research Agency in the framework of the “Investissements d’avenir” program (ANR-15-IDEX-02) and the “e-health and integrated care and trajectories medicine and MIAI artificial intelligence” Chairs of excellence from the Grenoble Alpes University Foundation. This work has been partially supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

The authors also acknowledge the Grenoble Alpes Data Institute, which is supported by the French National Research Agency under the “Investissements d’avenir” program (ANR-15-IDEX-02).

We thank Alison Foote for rereading of this article and improving our English.

DATA AVAILABILITY STATEMENT

Data availability is restricted.

ORCID

Guillaume Bottaz-Bosson https://orcid.org/0000-0001-6346-3510
Sébastien Bailly https://orcid.org/0000-0002-2179-4650

REFERENCES

1. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. 2019;7(8):687-698.

2. Lévy P, Kohler M, McNicholas WT, et al. Obstructive sleep apnoea syndrome. *Nat Rev Dis Primers*. 2015;1:15015.
3. Portier F, Frija EO, Chavaillon JM, et al. Traitement du SAHOS par ventilation en pression positive continue (PPC). *Rev Mal Respir*. 2010;27:S137-S145.
4. Siccoli MM, Pepperell JC, Kohler M, Craig SE, Davies RJ, Stradling JR. Effects of continuous positive airway pressure on quality of life in patients with moderate to severe obstructive sleep apnea: data from a randomized controlled trial. *Sleep*. 2008;31(11):1551-1558.
5. Pépin JL, Bailly S, Tamisier R. Big data in sleep apnoea: opportunities and challenges. *Respirology*. 2020;25(5):486-494.
6. Aardoom JJ, Loheide-Niesmann L, Ossebaard HC, Riper H. Effectiveness of electronic health interventions in improving treatment adherence for adults with obstructive sleep apnea: meta-analytic review. *J Med Internet Res*. 2020;22(2):e16972.
7. Genolini C, Ecochard R, Benghezal M, Driss T, Andrieu S, Subtil F. kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes. *PLoS One*. 2016;11(6):e0150738.
8. Gurrutxaga I, Muguerza J, Arbelaitz O, Pérez JM, Martín JI. Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recogn Lett*. 2011;32(3):505-515.
9. Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering – a decade review. *Inf Syst*. 2015;53:16-38.
10. Fréchet MM. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*. 1906;22(1):1-72.
11. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58(301):236-244.
12. Batagelj V. Generalized ward and related clustering problems. *Classification and Related Methods of Data Analysis*. Amsterdam, Netherlands: North-Holland; 1988.
13. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J Classif*. 2014;31(3):274-295.
14. Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S. Understanding and enhancement of internal clustering validation measures. *IEEE Trans Cybern*. 2013;43(3):982-994.
15. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Theory Methods*. 1974;3(1):1-27.
16. Gurrutxaga I, Albisua I, Arbelaitz O, et al. SEP/COP: an efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recogn*. 2010;43(10):3364-3373.
17. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224-227.
18. Kim M, Ramakrishna R. New indices for cluster validity assessment. *Pattern Recogn Lett*. 2005;26(15):2353-2363.
19. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern*. 1973;3(3):32-57.
20. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65.
21. Petitjean F, Ketterlin A, Gançarski P. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recogn*. 2011;44(3):678-693.
22. Sardá-Espinosa A. Comparing time-series clustering algorithms in R using the dtwclust package. Technical report; 2018.
23. Babbin SF, Velicer WF, Aloia MS, Kushida CA. Identifying longitudinal patterns for individuals and subgroups: an example with adherence to treatment for obstructive sleep apnea. *Multivar Behav Res*. 2015;50(1):91-108.
24. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2(1):193-218.
25. Lee IM, Shiroma EJ. Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. *Br J Sports Med*. 2014;48(3):197-201.
26. Fatouhi DE, Delrieu L, Goetzinger C, et al. Associations of physical activity level and variability with 6-month weight change among 26,935 users of connected devices: observational real-life study. *JMIR mHealth uHealth*. 2021;9(4):e25385.
27. Randerath W, Bassetti CL, Bonsignore MR, et al. Challenges and perspectives in obstructive sleep apnoea. *Eur Respir J*. 2018;52(3):1702616.
28. Pack AI. Application of personalized, predictive, preventative, and participatory (P4) medicine to obstructive sleep apnea. a roadmap for improving care? *Ann Am Thorac Soc*. 2016;13(9):1456-1467.

How to cite this article: Bottaz-Bosson G, Hamon A, Pépin J-L, Bailly S, Samson A. Continuous positive airway pressure adherence trajectories in sleep apnea: Clustering with summed discrete Fréchet and dynamic time warping dissimilarities. *Statistics in Medicine*. 2021;40:5373-5396. https://doi.org/10.1002/sim.9130

APPENDIX A. REAL CPAP DATA: DISTRIBUTION OF INDIVIDUAL ADHERENCE CHARACTERISTICS

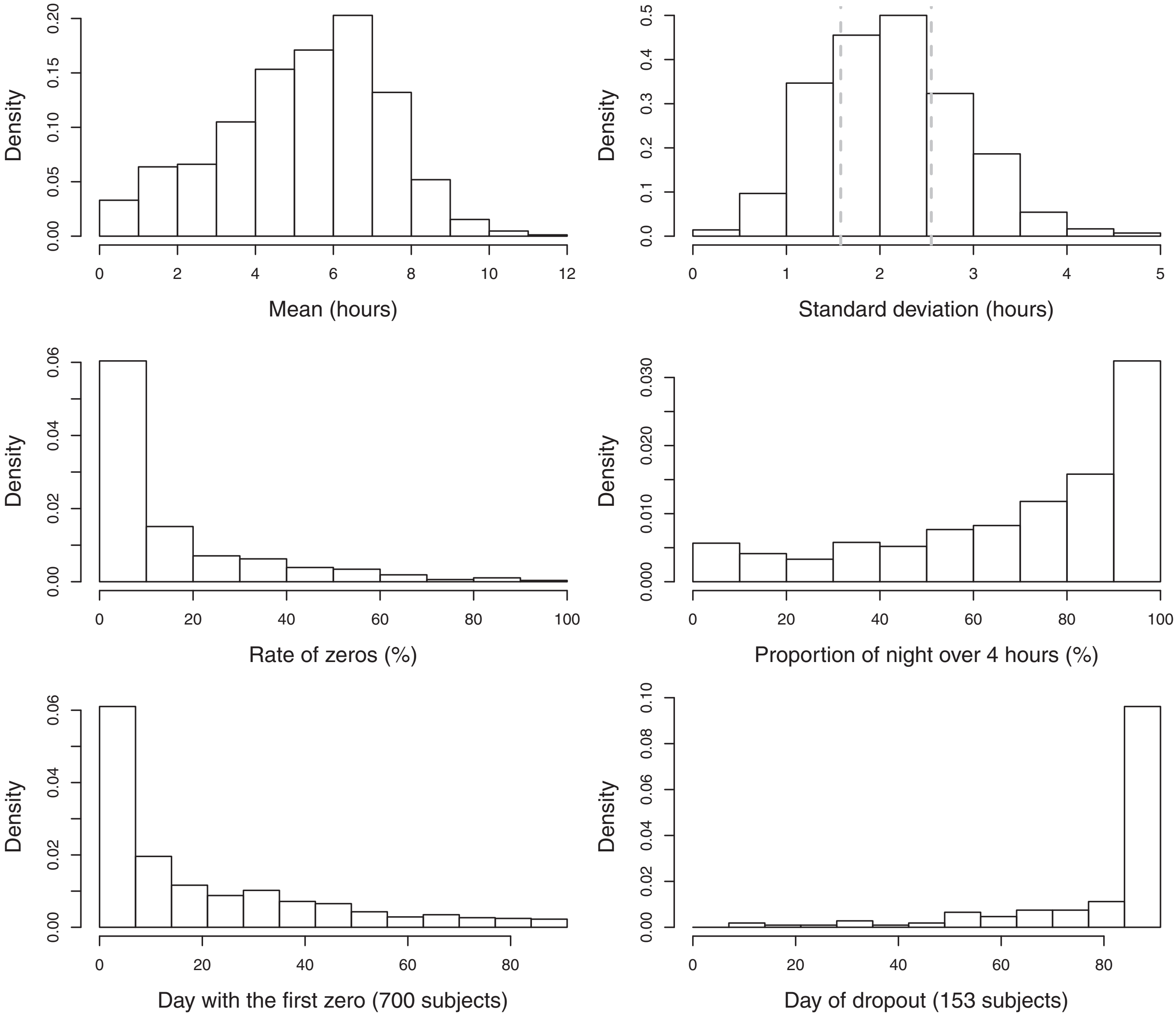


FIGURE A1 Distribution of individual adherence characteristics computed from the 91 days of observations: Mean (top-left), standard deviation (top-right), rate of null adherence (middle-left), proportion of nights with adherence over 4 hours (middle-right), day with the first zero (bottom-left), and day of dropout (bottom-right). The two dashed lines on the standard deviation plot represent the selected values used to generate the fictive datasets (with variance equal to 2.5 or 6.5)

APPENDIX B. MATHEMATICAL DEFINITIONS AND FORMULAS

B.1 Notations and preliminary definitions

Let $N \in \mathbb{N}^*$ the whole population size and $X := \{x_1, x_2, \dots, x_N\}$ be the whole dataset where x_i, $i \in [\![1, N]\!]$ is an individual's adherence time series. All time series are assumed to be of the same length $T \in \mathbb{N}^*$ and without missing data. For subject i, let $x_{i,t} \in \mathbb{R}$ represent her/his adherence value at the tth day of treatment, $t = 1, \dots, T$. So her/his adherence time series is $x_i = \left(x_{i,1}, x_{i,2}, \dots, x_{i,T}\right) \in \mathbb{R}^T$. For each time series x_i, let $\widetilde{x_i}$ be the polygonal curve specified by the ordered vertices sequence

$$\left(\widetilde{x_{i,1}}, \widetilde{x_{i,2}}, \dots, \widetilde{x_{i,T}}\right) = \left(\begin{pmatrix} 1 \\ x_{i,1} \end{pmatrix}, \begin{pmatrix} 2 \\ x_{i,2} \end{pmatrix}, \dots, \begin{pmatrix} T \\ x_{i,T} \end{pmatrix}\right) \in (\mathbb{R}^2)^T.$$

A *partition* P with k clusters on X is a set of non-empty subsets of X, $P = \{C_1, C_2, \dots, C_k\}$ such that $\bigcup_{C_l \in P} C_l = X$ and $\forall l, l',\ l \neq l' \Rightarrow C_l \cap C_{l'} = \emptyset$.

For a chosen dissimilarity measure d, and two time series x and x', the dissimilarity between x and x' is expressed as $d(x, x')$ and the dissimilarity between two clusters C_l and $C_{l'}$ computed through a linkage strategy is $d(C_l, C_{l'})$.

Then, let us define a warping path and a coupling, that are both included in the definitions of DTW and sdF dissimilarities.

Definition 1. A *warping path* W of lengths m and n is a sequence $\left((a_1, b_1), (a_2, b_2), \dots, (a_{l_W}, b_{l_W})\right)$ of distinct pairs from $[\![1, m]\!] \times [\![1, n]\!]$ such that: l_W is an integer non inferior to $\max(m, n)$, $a_1 = b_1 = 1$, $a_{l_W} = m$, $b_{l_W} = n$ and $\forall l$ in $\{1, \dots, l_W - 1\}$, $(a_{l+1} - a_l, b_{l+1} - b_l) \in \{(0, 1), (1, 0), (1, 1)\}$.

Definition 2. Let $E = (e_1, e_2, \dots, e_m)$ and $F = (f_1, f_2, \dots, f_n)$ be two sets of ordered elements of size m and n respectively. Let $W = \left((a_1, b_1), (a_2, b_2), \dots, (a_{l_W}, b_{l_W})\right)$ be a warping path of lengths m and n.

The sequence $L_W = \left((e_{a_1}, f_{b_1}), (e_{a_2}, f_{b_2}), \dots, (e_{a_{l_W}}, f_{b_{l_W}})\right)$ is called a *coupling* between E and F.

B.2 Dissimilarity measures

Definition 3. Let x_i and x_j be two times series. The *Euclidean distance* δ_E between x_i and x_j is defined by:

$$\delta_E\left(x_i, x_j\right) = \left\|x_i - x_j\right\|_2 = \sqrt{\sum_{t=1}^{T}\left(x_{i,t} - x_{j,t}\right)^2}.$$

Definition 4. Let x_i and x_j be two time series and $\mathcal{L}_{i,j}$ be the set of all possible couplings between x_i and x_j. The *dynamic time warping dissimilarity* d_{DTW} between x_i and x_j is defined by:

$$d_{DTW}\left(x_i, x_j\right) = \min_{L_W \in \mathcal{L}_{i,j}} \sum_{l=1}^{l_W} \left\|x_{i,a_l} - x_{j,b_l}\right\|_2.$$

Definition 5. Let $\widetilde{x_i}$ and $\widetilde{x_j}$ be the polygonal curves associated with the time series x_i and x_j and $\mathcal{L}_{i,j}$ be the set of all possible couplings between $\widetilde{x_i}$ and $\widetilde{x_j}$. Let $\lambda \in \mathbb{R}^+$ and $\Lambda : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\Lambda(a, b) = (\lambda a, b)$. The *generalized discrete Fréchet distance of parameter* λ, d_{dF_λ} between $\widetilde{x_i}$ and $\widetilde{x_j}$ is defined by:

$$d_{dF_\lambda}\left(\widetilde{x_i}, \widetilde{x_j}\right) = \min_{L_W \in \mathcal{L}_{i,j}} \max_{l=1}^{l_W} \left\|\Lambda(a_l, x_{i,a_l}) - \Lambda(b_l, x_{j,b_l})\right\|_2 = \min_{L_W \in \mathcal{L}_{i,j}} \max_{l=1}^{l_W} \left\|\Lambda(\widetilde{x_{i,a_l}}) - \Lambda(\widetilde{x_{j,b_l}})\right\|_2.$$

Thus the *generalized summed discrete Fréchet dissimilarity of parameter* λ, d_{sdF_λ} between $\widetilde{x_i}$ and $\widetilde{x_j}$ is defined by:

$$d_{sdF_\lambda}\left(\widetilde{x_i}, \widetilde{x_j}\right) = \min_{L_W \in \mathcal{L}_{i,j}} \sum_{l=1}^{l_W} \left\|\Lambda(\widetilde{x_{i,a_l}}) - \Lambda(\widetilde{x_{j,b_l}})\right\|_2.$$

B.3 Linkage strategies

The "average" linkage strategy defines the dissimilarity between two clusters C_l and $C_{l'}$ as:

$$d(C_l, C_{l'}) = \frac{1}{|C_l| \times |C_{l'}|} \sum_{x \in C_l} \sum_{x' \in C_{l'}} d(x, x').$$

The "complete" linkage strategy defines the dissimilarity between two clusters C_l and $C_{l'}$ as:

$$d(C_l, C_{l'}) = \max_{(x,x')\in(C_l \times C_{l'})} d(x, x').$$

The "Ward" linkage strategy recursively defines the dissimilarity between a merged cluster $C_{l'} \cup C_{l''}$ and a third cluster C_l as:

$$d^2(C_{l'} \cup C_{l''}, C_l) = \frac{|C_{l'}| + |C_l|}{|C_{l'}| + |C_{l''}| + |C_l|} d^2(C_{l'}, C_l) + \frac{|C_{l''}| + |C_l|}{|C_{l'}| + |C_{l''}| + |C_l|} d^2(C_{l''}, C_l) - \frac{|C_l|}{|C_{l'}| + |C_{l''}| + |C_l|} d^2(C_{l'}, C_{l''}),$$

where |.| denotes the number of elements of a cluster.

If two clusters are singletons then the dissimilarity value between these two clusters is set with the dissimilarity value between the corresponding objects.

APPENDIX C. SIMULATION STUDY: DISTRIBUTION OF NUMBER OF CLUSTERS RETURNED FOR EACH METHOD

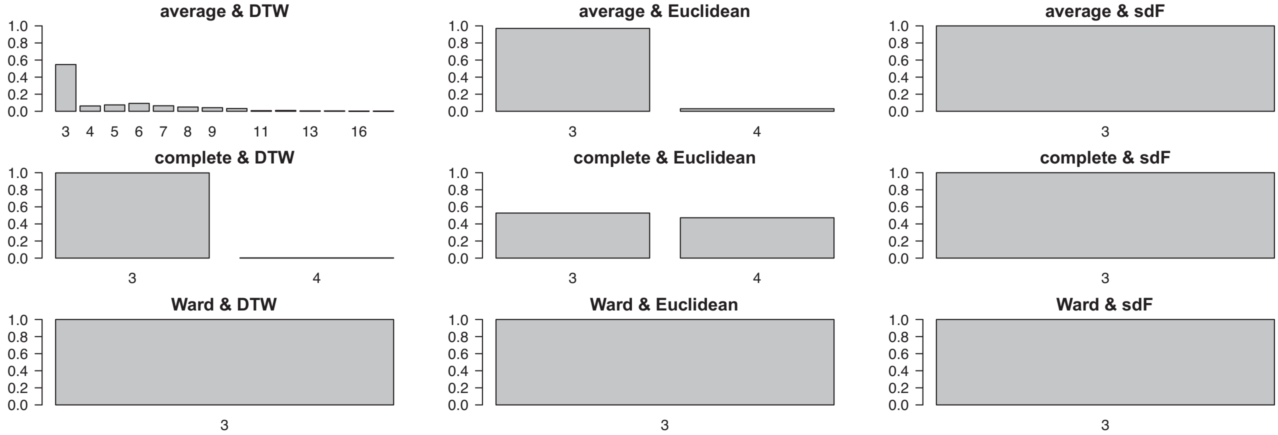


FIGURE C1 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 3 groups and without outliers

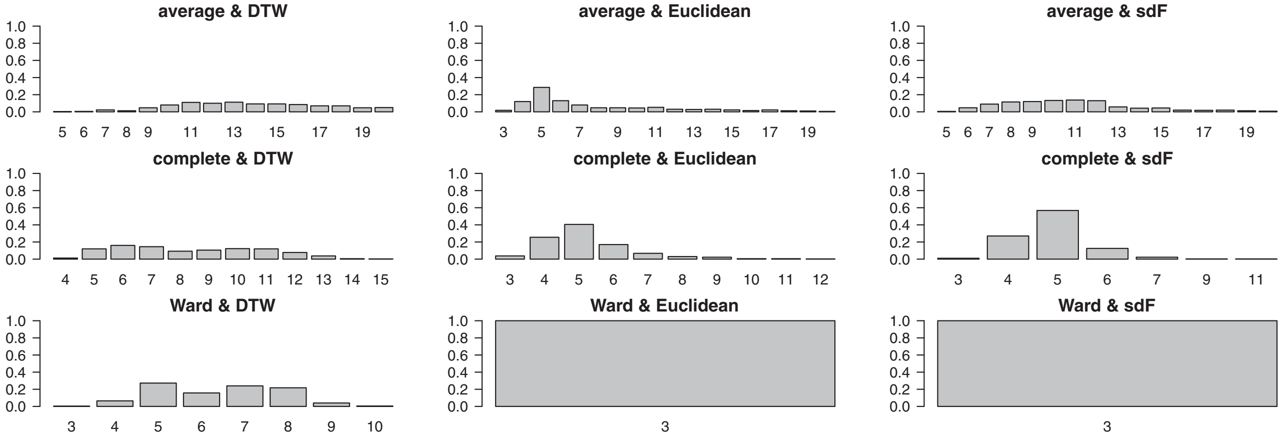


FIGURE C2 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 3 groups and 10% of outliers

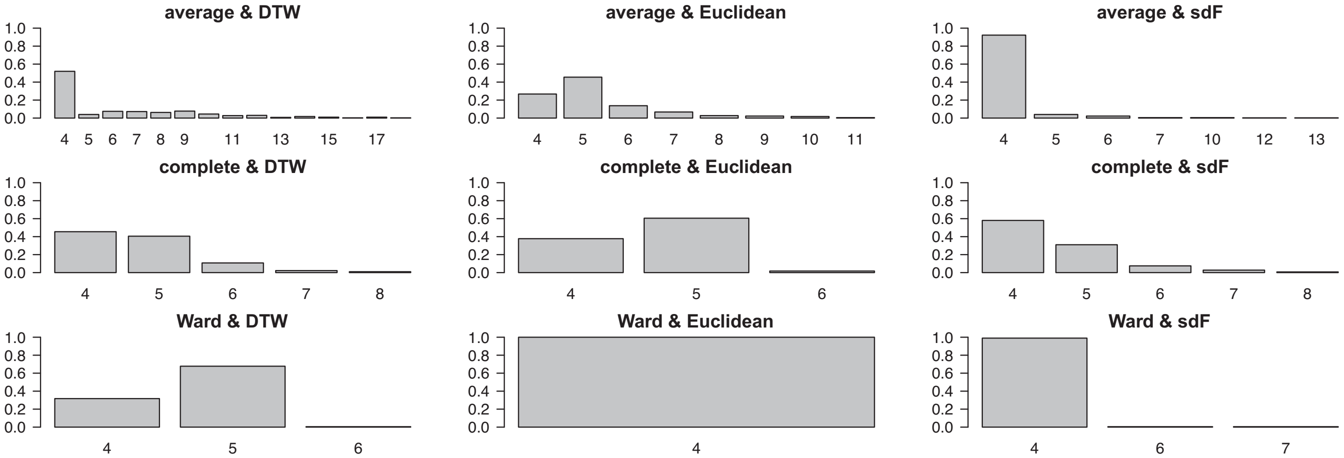


FIGURE C3 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 5 groups and without outliers

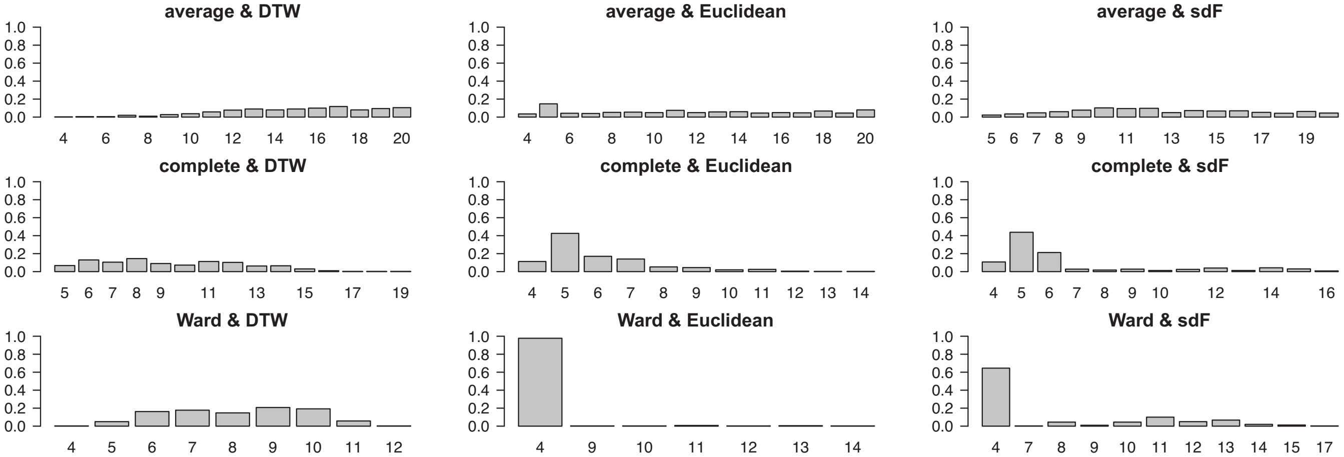


FIGURE C4 For each clustering method, distribution of simulated samples by the number of clusters that maximizes the ARI score: For the 400 samples with 5 groups and 10% of outliers

APPENDIX D. REAL CPAP DATA: MEDOID REPRESENTATIVES OF THE 6 CLUSTERS

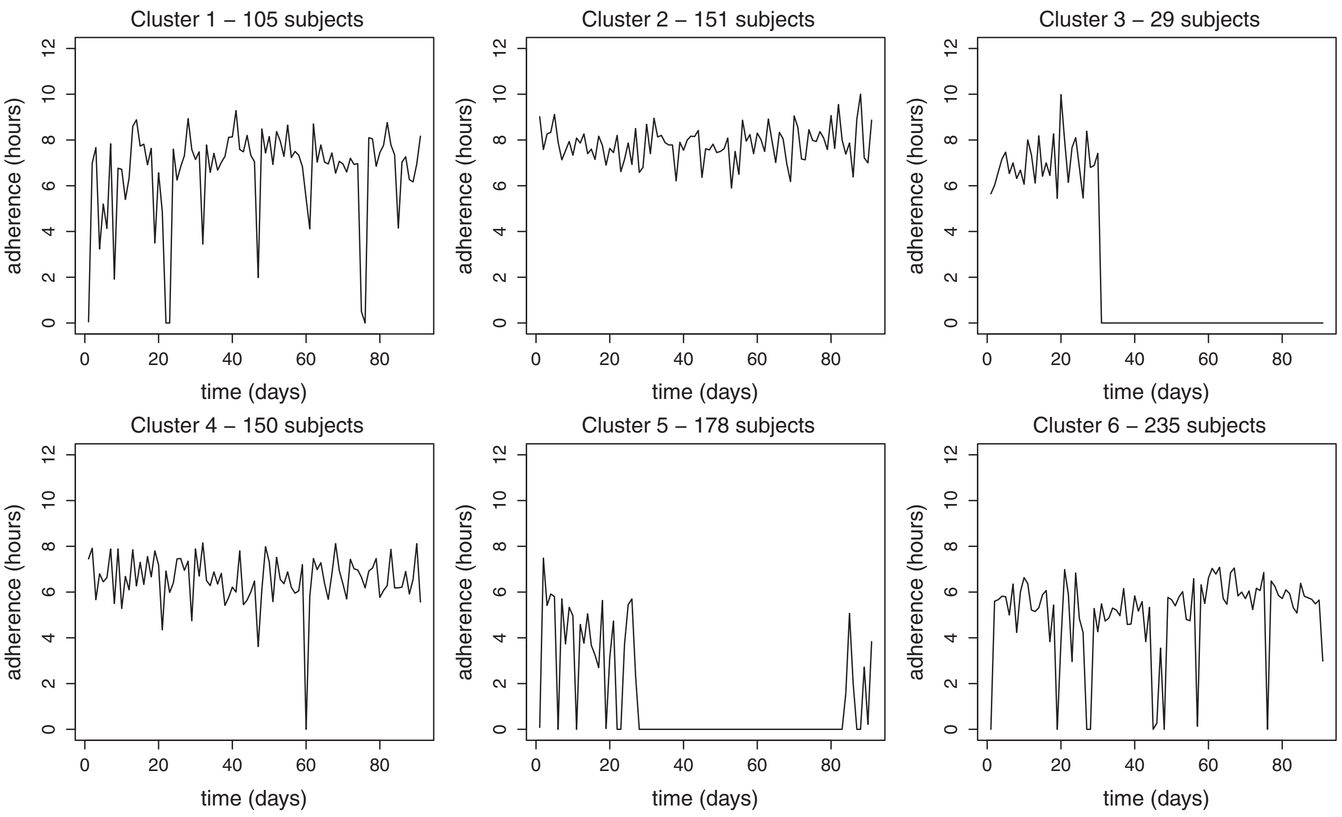


FIGURE D1 Real CPAP data: Medoid representatives of the 6 clusters