# Clustering Mixed Data Comprising Time Series

Loup-Noé Levy*
Energisme
92100 Boulogne-Billancourt, France
loup-noe.levy@energisme.com

Guillaume Guerard
Léonard de Vinci Pôle Universitaire
92 916 Paris La Défense, France
LI-PARAD Laboratory EA 7432
78035 Versailles, France
guillaume.guerard@devinci.fr

Soufian Ben Amor
LI-PARAD Laboratory EA 7432
78035 Versailles, France
soufian.ben-amor@uvsq.fr

Sonia Djebali
Léonard de Vinci Pôle Universitaire
92 916 Paris La Défense, France
sonia.djebali@devinci.fr

Clément Cornet
Léonard de Vinci Pôle Universitaire
92 916 Paris La Défense, France
clement.cornet@edu.devinci.fr

Maxence Choufa
Léonard de Vinci Pôle Universitaire
92 916 Paris La Défense, France
maxence.choufa@edu.devinci.fr

## ABSTRACT

The health and medicine sector is currently experiencing significant transformations, such as the integration of artificial intelligence in the decision-making process. In this complex system, there is a continuous data flow consisting of quantitative, qualitative, ordinal types, and time series. Hierarchical clustering is a useful tool to handle this complexity. However, clustering mixed data containing time series without distorting the inherent nature of the data poses a challenge. Although there are existing clustering techniques for mixed data or time series, the literature does not address the clustering of mixed data and time series. This paper presents several methodologies for such data clustering, including a novel algorithm based on pretopology. This hierarchical algorithm allows for customizable logical clustering, enabling health experts to better interpret and utilize the results for classification and recommendation by analyzing the hierarchy of clusters.

## CCS CONCEPTS

• **Information systems applications** → **Data mining**; Clustering; • **Probability and statistics** → Dimensionality reduction; • **Computing methodologies** → Machine learning.

## KEYWORDS

mixed clustering, time series clustering, pretopology

**ACM Reference Format:**
Loup-Noé Levy, Guillaume Guerard, Soufian Ben Amor, Sonia Djebali, Clément Cornet, and Maxence Choufa. 2018. Clustering Mixed Data Comprising Time Series. In *Proceedings of (SoICT 2023).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.


*SoICT 2023, December 7–8, 2023, Ho Chi Minh, Vietnam*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The popularity of mixed data clustering algorithms has increased due to the prevalence of real-world datasets containing both numeric and categorical features. Various methods have been proposed for clustering mixed data, though a unified research framework is still lacking in this field [3]. Researchers have used various types of clustering approaches for mixed data for heart disease [2], digital mammograms [13], acute inflammations [26], dermatology [28], cancer samples [35] and autism spectrum [32] for examples.

Time series features have also been an active area of research, with various methods proposed to address the challenges associated with handling time series. However, there is little to no scientific literature addressing the clustering of elements defined by categorical, numerical, and time series features, despite the presence of such data in various fields studying complex systems. For example, a patient in a hospital will not be described solely by fixed features, nor will they be described by time series alone. Fixed features are essential to cluster complex systems; for humans, it can be the date of birth, place of birth, or blood type, medical history, genetic disease, or disease predisposition.

Many features describing humans are time series, such as weight, blood glucose, blood pressure, and heart rate. In many contexts, these features are essential to administer a diagnosis. Since those data are often transformed into numerical values, increasing the curse of dimensionality, we propose various processes to cluster data involving fixed and fluctuating features, as it is necessary to identify homogeneous groups of complex elements.

Another challenge tackled in this article is the explainability, exploitability, and parametrization of heterogeneous and complex system clustering. Since unsupervised methods identify clusters in data without predefined labels, no clustering is inherently considered as the 'true' clustering. Ideally, the number of clusters, where they separate, and how depends on the specific needs and vision of each user and their context.

Hierarchical clustering is useful for handling this complexity, as it allows the user to identify coherent structures within each cluster, providing scalability and interpretability to the clustering. Allowing the user to adjust several clustering method parameters and easily understand their role in constructing clusters is also a way to enable more parametrization and interpretability of the clusters.

In this article, we will present the state of the art on mixed data clustering and time series clustering in Section 2. We will then present different possible strategies for clustering datasets composed of fixed and evolving features in Section 3. In this section, we also present a method based on the theory of pretopology that allows for hierarchical clustering of systems defined by mixed fixed and evolving features, and that allows for high parametrization, including a cluster construction based on logical rules defining the role of each feature in the clustering. The Section 4 shows the results for each methods and a discussion for future challenges. The Section 5 concludes the paper.

## 2 LITERATURE REVIEW

In the existing scientific literature, there is minimal focus on clustering data composed of numerical features, categorical features, and time series. However, there is substantial literature on mixed data clustering and time series clustering. In this section, we present relevant concepts and state-of-the-art methods for clustering and cluster evaluation of mixed data and time series.

### 2.1 Curse of dimensionality

The curse of dimensionality is a phenomenon that arises in high-dimensional spaces, particularly in clustering and machine learning tasks, where the increase in dimensions leads to exponentially larger search spaces, making it difficult for algorithms to operate efficiently [4, 33]. Furthermore, distance metrics that work well in lower-dimensional spaces may not be as effective in higher-dimensional spaces, leading to poor performance in clustering tasks [31]. This problem is particularly relevant in the context of complex data containing time series, as time series often have high dimensionality due to the numerous time points involved [33]. A solution to break this curse is often dimensionality reduction (DR).

DR is often employed as a preliminary step for clustering high-dimensional data. It can also be used to reduce mixed data. Factor Analysis of Mixed Data (FAMD) [10] is a DR technique specifically designed for such tasks. Additionally, although not initially adapted for mixed data reduction, Uniform Manifold Approximation and Projection (UMAP) [21] or Pairwise Controlled Manifold Approximation Projection (PaCMAP) [34] can be adapted. UMAP is adapted by using the Huang Distance, that is suited for mixed data, and PaCMAP can be initialized with FAMD. Then, these techniques are able to convert a high-dimensional mixed dataset into a low-dimensional numerical dataset. Subsequently, state-of-the-art numerical clustering algorithms, such as K-means, can be applied to the transformed dataset, and cluster visualization on the low-dimensional data can be performed. DR is also a prevalent preprocessing approach for time series clustering, aiming to decrease the complexity and computational cost associated with high-dimensional data.

### 2.2 State of the Art in Mixed Data Clustering

Let us introduce several well-known methods for mixed data clustering.

*Partitioning Clustering for Mixed Data.* aims to divide a dataset into a predetermined number of non-overlapping clusters. **K-Prototypes** [15] extends the traditional K-Means algorithm to handle both numerical and categorical features by combining K-Means for numerical data and K-Modes [14] for categorical data. **Convex K-Means** [22] iteratively refines the centroids of $K$ clusters (with a convex hull) until convergence or a maximum number of iterations is reached. These methods provide effective clustering solutions for mixed data.

*Model-Based Clustering for Mixed Data.* utilizes statistical models to describe the distribution of data points within each cluster, accommodating numerical, categorical, and time series features. **MixtComp** [5] is a statistical method that combines the strengths of model-based clustering and Bayesian approaches, handling different types of data and missing data. **KAMILA** [11] is a clustering algorithm designed to handle mixed data by extending the standard K-means algorithm, combining K-means clustering with the Gaussian-multinomial mixture model. These methods provide effective clustering solutions for mixed data.

*Hierarchical Clustering for Mixed Data.* enables the exploration of subgroups that constitute a cluster. **Philip and Ottaway** [27] proposed a method based on Gower's similarity measure, which effectively computes the similarity between mixed data points for hierarchical clustering. **DenseClus**[1] is a density-based hierarchical clustering algorithm that performs DR using UMAP and employs the accelerated HDBSCAN algorithm [20]. **Agglomerative Hierarchical Clustering (AHC)** for mixed data is a powerful technique that effectively groups data points with different feature types based on parameterised distance metrics [8]. These methods enable meaningful clustering of mixed data by accounting for differences between categorical and numerical features.

### 2.3 State of the Art on Time Series Clustering

Time series clustering has been an area of active research for several years due to its widespread applicability in fields such as finance, healthcare, and IoT [1]. The primary goal of time series clustering is to group similar time series, considering their temporal dynamics and patterns. This section reviews the state of the art in time series clustering.

*Distance-based Clustering.* is one of the most common approaches for clustering time series. This approach computes pairwise distances between time series, using a distance metric to measure similarity. The most widely used distance metrics for time series clustering are Euclidean distance, **Dynamic Time Warping (DTW)** [23], and **Longest Common Subsequence (LCSS)** [9]. DTW is particularly popular because it allows for non-linear alignment between time series, providing a more flexible similarity measure compared to the Euclidean distance.

*Feature-based Clustering.* involves extracting **time series features (TSF)** and using these features to represent the time series in a lower-dimensional space. This approach can reduce the dimensionality of the data and the computational complexity associated with clustering. Common features extracted from time series include statistical features (e.g., mean, standard deviation), frequency domain features (e.g., Fourier transform, wavelet transform), and shape-based features [12].

[1] https://github.com/awslabs/amazon-denseclus

*Multivariate time series feature extraction.* involves deriving additional features or new time series from the analysis of links between two or more time series. Specific features can be extracted depending on the case study; for instance, in building classification and clustering, features are often calculated based on outside temperature and energy consumption. In general, the extracted features involve the evaluation of the correlation between different time series [30]. The use of autoregressive modeling to form augmented-feature vectors [16] is also an option. After feature extraction, traditional clustering algorithms, such as k-means or hierarchical clustering, can be applied to the reduced feature space.

*Model-based Clustering.* methods assume that each time series is generated by an underlying model, and the goal is to group time series based on the similarity of their generative models. Some popular model-based clustering techniques include clustering based on **Hidden Markov Models (HMM)** [24], **autoregressive models** [18], and **Gaussian Process models** [19]. These methods often require fitting models to each time series and comparing the models to compute pairwise similarities, which can be computationally expensive.

## 2.4 Cluster evaluation

Evaluating the quality of clusters is more challenging than evaluating classifications due to the absence of ground truth for comparison. Instead, the focus is on the quality of a partition, based on metrics such as dispersion and distances within and between clusters [25].

*Calinski-Harabasz (CH).* The CH index [6] is a well-established metric for evaluating the definition of clusters. The index, also known as the Variance-Ratio Criterion, is computed as the ratio of the sum of between-cluster dispersion and within-cluster dispersion for all clusters, where dispersion is the sum of squared distances. A clustering with a high CH score indicates a model with well-defined clusters. This method provides a robust approach to assess the quality and explainability of mixed data clustering. A higher CH score is indicative of a model that exhibits more distinct and well-defined clusters.

*Davies-Bouldin (DB).* The DB index [7] is used to assess the separation of clusters. The similarity between a pair of clusters is the ratio of the sum of the average distance in each of the two clusters and the distance between the centroids of the two clusters. The DB index is then computed as the average of the maximum similarities for each individual cluster. Lower values of the DB index signify better-separated clusters, with the minimum possible score being zero.

*Silhouette Coefficient (SC).* The SC, first presented in [29], indicates how well-defined the clusters are. A score is calculated for each data point as shown in Equation 1, where $w$ represents the average distance of a point to other points within its cluster, and $c$ represents the average distance to points in the closest cluster.

$$SC = \frac{c - w}{max(w, c)} \tag{1}$$

The SC for a group of points is determined by averaging the SC of each individual sample. The coefficient ranges between -1 for improper clustering and +1 for highly compact clustering. A score of zero implies that clusters are overlapping.

One issue with mixed data clustering is that these metrics are defined for numerical spaces. Therefore, the application of the DR techniques described earlier is necessary for any cluster evaluation. Similarly, for complex data, it must be reduced before the clusters are evaluated. In order to calculate the CH, DB, and SC scores, datasets are transformed into Euclidean spaces using FAMD, which ensures that the output space has the same number of dimensions as the original space. FAMD is chosen for its known inertia, deterministic nature, and minimal reliance on hyper-parameters. Additionally, since the SC score is the only index in the study that accepts a pairwise distance matrix as input, it is computed using the Gower matrix to prevent any bias towards FAMD or provide extra insights when FAMD has low inertia. We will call it the **Gower Silhouette Coefficient (GSC)**.

## 2.5 Pretopology-based clustering PretopoMD

This subsection introduces the essential concepts and definitions in pretopology, such as pretopological space and pseudoclosure, before describing the primary algorithm for pretopological hierarchical clustering.

A pseudoclosure function, denoted as $a : \wp(U) \rightarrow \wp(U)$, operates on a set $U$ of elements and adheres to the conditions $a(\emptyset) = \emptyset$ and $\forall A \mid A \subseteq U : A \subseteq a(A)$, where $\wp(U)$ symbolizes the power set of U. A pretopological space consists of a tuple $(U, a(.))$, with $U$ representing a set of elements and $a(.)$ being a pseudoclosure function on $U$. In this space, closure is determined by iteratively applying the pseudoclosure operator to the set and its subsequent images until no further expansion occurs. The closure of a subset $A$ of $U$ is the smallest closure containing $A$, represented as $F(A)$. The process of constructing a hierarchy involves recursive pseudoclosure steps that connect any set of elements to a larger set.

The framework for formalizing a pretopological space, adapted from Julio Laborde's work [17], characterizes a pretopological space with a tuple $(G, \Theta, DNF(.))$. Here, $G$ denotes a collection of $n$ weighted directed graphs, $\Theta$ signifies a set of $n$ thresholds associated with each graph, and $DNF(.)$ represents a boolean function defined as a positive **Disjunctive Normal Form (DNF)** involving $n$ boolean functions $V_1(A, x), ..., V_n(A, x)$, each associated with a graph. The truth value depends on the set $A$ and element $x$.

To determine if an element $x \in U$ belongs to the pseudoclosure of a set $A$, follow these steps: For each $V_i(A, x)$, $V_i(A, x) = True$ if and only if $\sum_{e_{xy} \in G_i, y \in A} w(e_{xy}) \geq \theta_i$, where $e_{xy}$ denotes the edge from $x$ to $y$, and $w(e)$ represents the edge weight $e$. The element $x \in U$ belongs to the pseudoclosure of $A$ if and only if the $DNF(.)$ evaluates to True.

In summary, this process checks if the sum of edge weights connecting element $x$ to the elements within $A$ is greater than the threshold associated with the graph in each graph. If this condition is satisfied, the boolean variable corresponding to that graph takes the value True; otherwise, it takes the value False. The element belongs to the pseudoclosure if $DNF(.)$ evaluates to True given the values of the boolean functions $V_i(A, x)$.

The primary insight obtained from this pretopological framework and its associated algorithm is that pretopology enables the

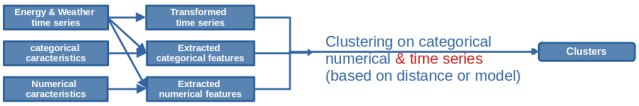


**Figure 1: Clustering on Mixed Features and Time Series based on distance or model.**

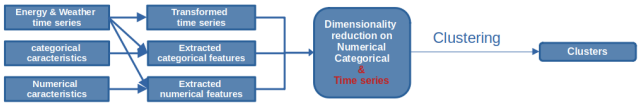


**Figure 2: Clustering on numerical data only via DR.**

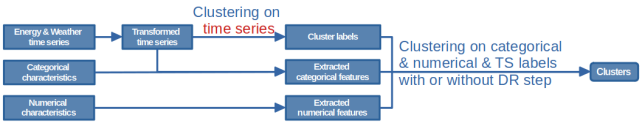


**Figure 3: Clustering on mixed features and pre-clustered Time Series labels as categorical features.**

abstraction of the complex nature of the elements being studied by focusing on the relationships between them based on their characteristics. Each characteristic has its own weighted graph, which allows the calculation of a distance for each characteristic. For example, a Manhattan distance can be computed for a pair of longitude and latitude coordinates, while a corresponding volume difference can be calculated for a 3D space describing an object's dimensions. Similarly, the distance between two highly time-dependent series can be measured using Euclidean space, while Dynamic Time Warping (DTW) can be employed to compare time series where the overall profile is more relevant. Once this set of graphs is defined, the Disjunctive Normal Form (DNF) establishes the logical rules by which pseudoclosure, and consequently hierarchical clustering, are generated.

## 3 METHODS

In this section, we will introduce various approaches that appear relevant for clustering complex data. These approaches are combinations of the different components presented in the state of the art.

*Method1: Clustering on Mixed Features and Time Series using each time step as dimension.* (Figure 1)

This approach involves using each time step of the time series as a numerical feature and applying mixed clustering methods such as K-prototype.

**Advantages:** Simple to implement and use state-of-the-art mixed clustering methods.

**Disadvantages:** In this case, the "weight" of each measure of the time series is the same as other features, and simple numerical or categorical features will be overshadowed due to the sheer volume of time series values, leading to inadequate consideration in the resulting clusters.

**XAI:** Medium.

*Method2: Clustering on Mixed Features and Time Series using specific distances.* (figure 1)

In this approach, specific distances are calculated for particular features or groups of features. These distances are subsequently aggregated in the clustering process, using either a weighted sum or logical rules.

**Advantages:** All available information is fully exploited to create the most relevant clusters possible.

**Disadvantages:** This approach requires a deep understanding of the dataset and specification of the appropriate AHC or PretopoMD.

**XAI:** High to very high.

*Method3: Clustering on numerical data only via Dimensionality Reduction.* (figure 2)

In this method, we use dimensionality reduction (DR) to create a low-dimensional numerical representation of all features (numerical, categorical, and time series). To apply DR on time series, we consider a time series as a point in a high-dimensional space, where each time step is a dimension.

**Advantages:** State-of-the-art numerical clustering methods can be applied.

**Disadvantages:** DR of long time series can be challenging due to the "curse of dimensionality" (see Section 2).

**XAI:** Very low.

*Method4: Clustering on mixed features and pre-clustered Time Series labels as categorical features.* (figure 3)

In this method, we apply one or several time series clustering methods on each time series. We obtain a label corresponding to which cluster the time series belongs to. This label is then considered as a categorical feature. Mixed data clustering methods are then used on the enriched dataset.

**Advantages:** The similarity between time series is considered through the time series clustering methods. State-of-the-art mixed dataset clustering methods can be applied.

**Disadvantages:** The choice of methods (including metrics and hyperparameters) can affect the quality of time series clusters labels.

**XAI:** Very high.

*Method5: Clustering on mixed features and pre-clustered time series labels as categorical features using DR.* (figure 3)

In this method, the same preliminary steps as in method 4 are executed to obtain time series labels. Next, a DR method is applied to create a numerical dataset on which numerical clustering is performed.

**Advantages:** State-of-the-art numerical clustering methods can be applied.

**Disadvantages:** DR can create a loss of information and explainability.

**XAI:** Low.

*Method6: Clustering on Mixed Features and Time Series Features.* (figure 4)

In this method, as in methods 4 and 5, we do not apply clustering on time series in their raw form. Instead, we extract TSFs to capture their essence in numerical and categorical representations. By doing

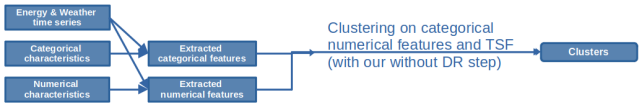


**Figure 4: Clustering on Mixed Features and Time Series Features**

so, we enable the application of mixed clustering methods on the extracted data.

**Advantages:** Mixed clustering methods can be applied. The features extracted are specific to the context and therefore can allow the clustering to be relevant from a field point of view

**Disadvantages:** The amount of information lost during preprocessing is important, though varying depending on the nature of the time series and the quality of the feature extraction process.

**XAI:** High.

*Method7: Clustering on Mixed Features and Time Series Features using DR.* (figure 4)

We add a dimensionality reduction step to the sixth method.

**Advantages:** State-of-the-art numerical clustering methods can be applied.

**Disadvantages:** Information loss both during feature extraction and DR

**XAI:** Low.

# 4 RESULTS AND DISCUSSION

## 4.1 Presentation of the Test Dataset

Since health datasets can be technically long to explain and to display, we present a generated dataset with categorical, numerical features and time series.

To evaluate the clustering methods, we generated a dataset consisting of elements characterized by four features: their position in a 2D space (numerical), their size (numerical), their shape (categorical with four possible values), and a time series consisting of a hundred data points. The dataset comprises 50 elements. The motivation for selecting such a dataset was to enable visualization without the need for DR, allowing for a direct understanding of the cluster construction (see Figures 5 and 6). This approach demonstrates how the logical rules defining the PretopoMD algorithm can enable customized clustering that addresses specific field requirements.

## 4.2 Cluster Quality Indicators

Using the evaluation metrics presented in Section 2.4, we assessed the outcomes of various clustering algorithms.

To calculate the CH, DB, and SC scores, we transformed datasets into Euclidean spaces using FAMD, ensuring that the output space has the same number of dimensions as the original space. We chose FAMD as the dimensionality reduction method because it is not too dependent on hyperparameters, because the inertia of the model is known (as it is a factorial method), and for its deterministic nature.

Moreover, since the SC score is the only index in this study that can accept a pairwise distance matrix as input, we also computed it using the Gower matrix. This may prevent any bias towards FAMD

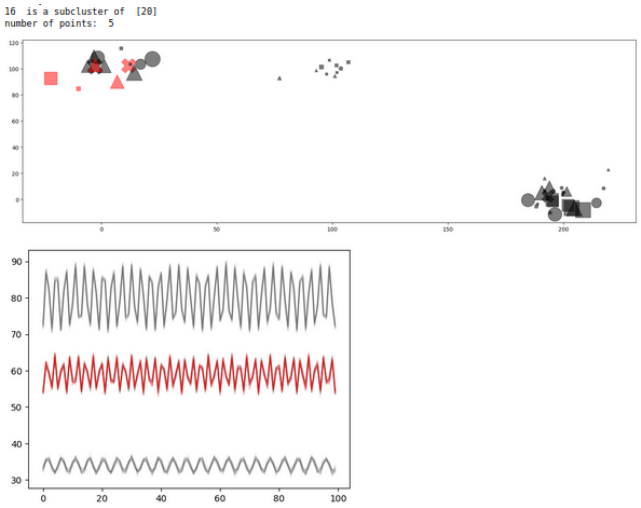


**Figure 5: In this example, the hierarchical clustering has been made using the DNF condition *Position AND TS*. Thus, the subcluster elements are spatially close and have similar time series.**

and provide additional insights in cases where FAMD achieves low inertia.

It should be noted that in certain situations, an algorithm might produce a single cluster or only outliers. In such cases, we present the worst possible score or an infeasible value. The results presented here are from a generated dataset (subsection 4.1, more results and information are available in the GitHub repository[2].

## 4.3 Clustering on Features and Time Series: Pretopological Clustering

Here we will present the results of PretopoMD. We have created four prenetworks for this instance: one for the position features, one for the size feature, one for the shapes of the elements, and one for the time series. For each prenetwork, a distance matrix is calculated. We use Euclidean distance for numerical features, Hamming distance for categorical features, and Dynamic Time Warping (DTW) for time series.

Different DNFs were used, and the DNF that scored highest on CH, SC, and GSC was the one using only TS, indicating that clustering based solely on the time series was more effective than using more complex clustering methods. The only score that did not favor this DNF was DB, which preferred *Position AND Size AND TS OR Shape*. This DNF provided better cluster separation because it had more *AND* rules, which made it more likely to divide the dataset into many clusters with similar position, size, and time series characteristics.

However, other DNF combinations could be chosen depending on the user's needs, as the relevance of the clustering varies according to the application. For illustrative purposes, a clustering using the simple *Position AND Time Series* rule is shown in Figures 5 and 6, identifying 8 clusters. Each cluster consists of elements

[2]https://github.com/Loup-Noe/clustering-mixed-data-comprising-time-series

| Method | CH | DB | SC | GSC |
|---|---|---|---|---|
| **Method 1** | | | | |
| AHC_Gow_3 | 8.01 | 2.69 | 0.12 | 0.93 |
| Kamila | 8.01 | 2.69 | 0.12 | 0.93 |
| K-Prototypes | 8.01 | 2.69 | 0.12 | 0.93 |
| PretopoMD_Euclid_Hamm | 4.27 | 2.16 | -0.04 | -0.26 |
| **Method 2 AHC** | | | | |
| AHC_Gow_DTW_6 | 3.61 | 4.64 | -0.03 | 0.38 |
| AHC_Gow_DTW_5 | 4.38 | 4.55 | 0.01 | 0.52 |
| AHC_Gow_DTW_4 | 5.79 | 3.07 | 0.06 | 0.73 |
| AHC_Gow_DTW_3 | 8.01 | 2.69 | 0.12 | 0.93 |
| **Method 2 PretopoMD** | | | | |
| Pos_&_Size_or_Shape_&_TS | 4.10 | 4.90 | 0.06 | 0.54 |
| Pos_or_Size_&_Shape_or_TS | 4.10 | 4.90 | 0.06 | 0.54 |
| Pos_&_Size_&_TS_or_Shape | 7.48 | 1.01 | 0.12 | 0.19 |
| Pos_&_Size_or_Shape_&_TS | 1.20 | 3.22 | -0.27 | -0.49 |
| Pos_&_TS | 2.58 | 3.75 | -0.14 | 0.28 |
| TS | 8.01 | 2.69 | 0.12 | 0.93 |
| **Method 3** | | | | |
| DenseClus | 0.00 | -1.00 | -1.00 | -1.00 |
| FAMD-KMeans | 26.84 | 1.03 | 0.48 | -0.07 |
| PretopoMD-FAMD | 28.06 | 0.81 | 0.51 | -0.07 |
| PretopoMD-Laplacian | 0.40 | 3.62 | -0.12 | -0.28 |
| PretopoMD-UMAP | 7.49 | 2.76 | 0.11 | 0.87 |
| PretopoMD-PaCMAP | 25.84 | 1.03 | 0.47 | -0.08 |
| PretopoMD-Louvain | 2.97 | 3.30 | -0.16 | 0.21 |
| **Method 4** | | | | |
| DenseClus | 0.00 | -1.00 | -1.00 | -1.00 |
| AHC_Gow_3 | 1.40 | 5.77 | 0.01 | 0.00 |
| K-Prototypes | 1.08 | 6.57 | 0.00 | 0.01 |
| PretopoMD_Eucl_Hamm | 1.80 | 2.49 | -0.25 | -0.51 |
| **Method 5** | | | | |
| FAMD-KMeans | 1.08 | 6.57 | 0.00 | 0.01 |
| PretopoMD-FAMD | 3.86 | 2.07 | -0.06 | -0.15 |
| PretopoMD-Laplacian | 1.05 | 5.81 | -0.18 | -0.34 |
| PretopoMD-UMAP | 0.95 | 7.38 | -0.09 | -0.12 |
| PretopoMD-PaCMAP | 11.04 | 1.79 | 0.23 | -0.04 |
| PretopoMD-Louvain | 1.08 | 5.32 | -0.05 | -0.15 |
| **Method 6** | | | | |
| AHC_Gow_3 | 8.43 | 2.51 | 0.11 | 0.57 |
| K-Prototypes | 7.67 | 2.51 | 0.08 | 0.38 |
| PretopoMD_Eucl_Hamm | 5.99 | 2.10 | 0.10 | 0.11 |
| **Method7** | | | | |
| DenseClus | 0.00 | -1.00 | -1.00 | -1.00 |
| FAMD-KMeans | 16.73 | 1.37 | 0.32 | 0.08 |
| PretopoMD-Laplacian | 2.21 | 3.66 | 0.02 | 0.01 |
| PretopoMD-UMAP | 6.34 | 2.65 | 0.09 | 0.49 |
| PretopoMD-PaCMAP | 15.79 | 1.40 | 0.32 | 0.06 |
| PretopoMD-Louvain | 6.45 | 2.21 | 0.06 | 0.42 |
| PretopoMD-FAMD | 2.48 | 2.36 | -0.03 | -0.13 |

**Table 1: Cluster evaluation scores of the different methods**

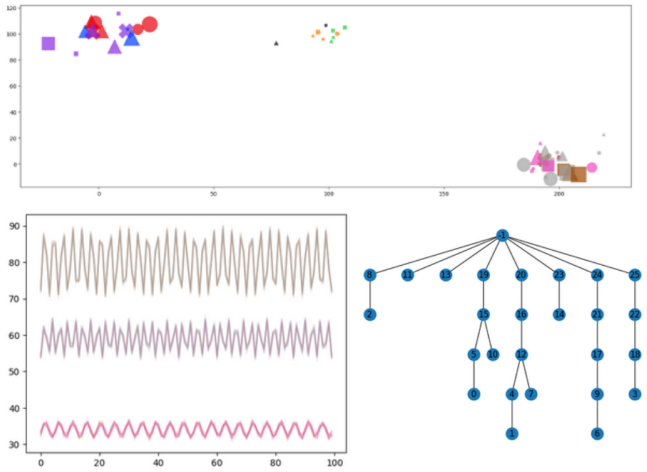

**Figure 6: Visualisation of mixed hierarchical clustering using time series.**

that are close in space and have similar time series patterns. This observation also applies to the subclusters within the larger clusters. The result of a more complex DNF, corresponding to more specific needs, is also presented in Figure 8.

## 4.4 Discussion

We observe that PretopoMD-FAMD achieves the best results in terms of CH, DB, and SC Scores in Table 1 of results. This can be attributed to the fact that clustering in conjunction with DR is highly effective on time series data as it mitigates the curse of dimensionality. It is also worth noting that the CH, DB, and SC Scores are all calculated on the dataset after applying FAMD, thereby favoring clustering methods that utilize FAMD in their preprocessing. Had we evaluated the clusters using CH, DB, and SC by reducing the dataset with another dimensionality reduction method, we would have obtained different results. Additionally, we can note that PretopoMD-FAMD does not have a good score on GSC despite being the best on the other metrics.

If we normalize and add up our scores, the best algorithms in descending order are: PretopoMD-FAMD, FAMD-KMeans, PretopoMD-PaCMAP, FAMD-KMeans with TSF instead of the whole time series, and PretopoMD-PaCMAP with TSF instead of the whole time series. Just below these are AHC_Gow_DTW with three clusters, Phillip and Ottaway, Kamila, K-Prototypes, and PretopoMD using only time series values. These methods have identified clusters that correspond exclusively to the time series.

Interestingly, these methods that identified only the time clusters (AHC with three clusters, Phillip and Ottaway, Kamila, K-Prototype, and PretopoMD using the DNF Time Series) achieved the highest GSC Score, all at equal values. It means that the other features are not only deemed irrelevant by these clustering methods but also the GSC Score. For example, AHC with more than three clusters employ other features for clustering but are considered worse than AHC_3.

There are several interpretations of this phenomenon. First, K-Prototypes, Kamila, and Phillip and Ottaway treat each time step

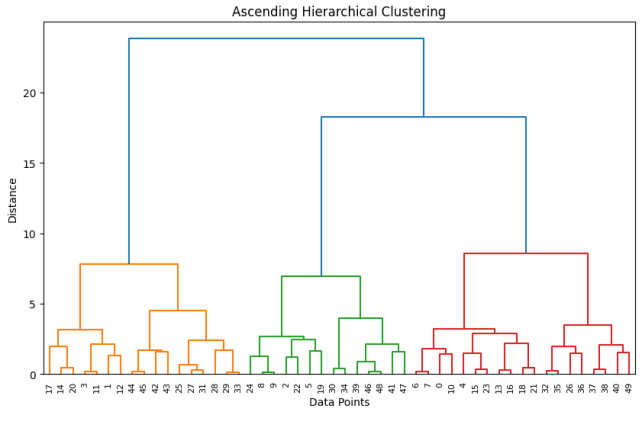


**Figure 7: The DTW distance between the time series slightly outweighs the Gower distance between the mixed features in this dataset. All evaluation metrics rewards the separation in 3 clusters**

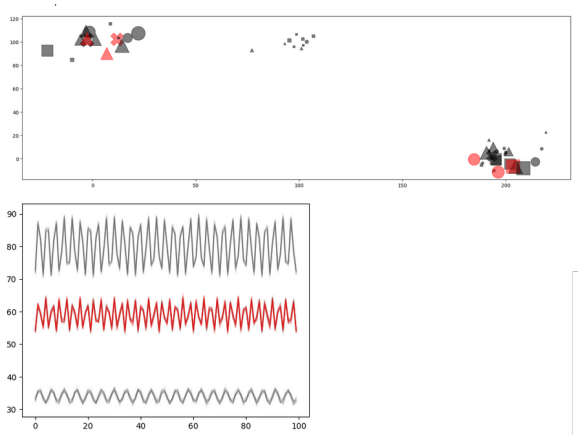

**Figure 8: A subcluster of the hierarchy build with the DNF *(Position AND Shape AND TS) OR (Size AND TS)* prioritizes *TS*, then *Size* then equally *Position* and *Shape***

as a feature, making non-time series features less significant in the resulting clusters due to their comparatively low numbers.

As for AHC, it was specifically designed to address this issue by incorporating a distance specific to time series in addition to the Gower distance for other features. By examining the dendrogram in Figure 7, we can observe that the three clusters are well-separated because the distance between time series is more pronounced than the distance between other features. However, this is more attributable to the test dataset, in which the time series are extremely similar, rather than the Hierarchical Clustering methods itself. In this instance, when clustering into three clusters, it made sense to cluster based on time series similarity. When more clusters were demanded from AHC, it provided a finer separation of the dataset, taking into account other features. However, no indicators rewarded such behavior. This is the case with and without normalized distance in AHC. What was not attempted here was assigning weights to different distances based on specific needs or characteristics. In a case study, one could decide to give more weight to a certain set of parameters for them to have a more significant influence on the resulting hierarchical clustering.

Another point concerning the evaluation metrics is that none of the extracted features used for some of the clustering were added to the dataset. Adding the dataset with pre-identified time series clusters or extracted features might have changed the way the clusters are evaluated.

## 4.5 Challenges and Future Works

Exploring further cluster evaluation metrics and aggregating them might be a solution for hyperparameterization. The objective might not simply be to have the highest average score but to find a clustering that has scores relatively high for all metrics.

However, one must accept that the quality of clustering is highly dependent on the objectives of the user, especially in the case of complex data. Depending on the case study, the relevance of one aspect of the data can vary significantly. Visually analyzing the data in its raw decomposed form, such as in time series, or visualizing it through different DR techniques can allow users to view it from various perspectives (quite literally) and realize how one clustering might seem more appropriate when viewed through FAMD and another more relevant when viewed through UMAP. Ultimately, it is the meaning behind the features and the coherence of the final clusters that give relevance to a clustering method.

Therefore, hierarchical clustering techniques such as PretopoMD, which function extremely well with DR, might actually be more relevant without DR when complex rules and distances must be used to identify clusters according to specific requirements. AHC might also be used in this manner simply through the use of weights. Both have the advantages of allowing the user to zoom in on a cluster to identify subgroups, which is often relevant in complex data contexts. For example, the AHC dendrogram allowed us to view how the relatively high distance between the time series cluster influenced the separation of the complex dataset and how weighting the different distance might have changed this separation (see figure 7).

Regarding pretopology, the example in figure 5, as well as more complex DNFs allow for some very interesting hierarchie. For example, a hierarchical clustering built with the DNF *(Position AND Shape AND TS) OR (Size AND TS)* (see figure 8) will return the same clusters as the DNF *TS*, but will return a hierarchy with subclusters of elements that are necessarily close in terms of time series but are also as close as possible in terms of position, size, or shape, with size being the first criterion of aggregation. That is, the smaller clusters are necessarily close in time series and are mostly close in size. Then they expand by integrating other elements according to the other criteria. Adjusting the DNF in this manner enables the construction of hierarchies tailored to meet the complex requirements specific to various case studies. Furthermore, besides the DNF, the diverse parameters of PretopoMD facilitate extensive customization of the dispersion, size, and number of outliers within the clusters.

An effective approach to address the explainability issues associated with clustering, particularly in the case of complex data, is to have a comprehensive understanding of both the dataset and the

various steps involved in a clustering method, while adhering to logical rules and parameter settings. Developing improved visualization tools that can help in understanding the value of clusters in high-dimensional contexts is also an important area of research.

In our future work, we plan to explore these solutions while working on a large and diverse dataset that includes energy consumption, weather time series, and building mixed characteristic typologies. This will allow us to focus on meaningful feature extraction and construct clusters in collaboration with field experts, by analyzing the significance of the clusters and adjusting parameters and logical rules accordingly. Additionally, we will investigate automated hyperparameter tuning based on a combination of quality indicators.

## 5 CONCLUSION

In conclusion, this paper addresses the challenge of clustering mixed data containing time series, which is prevalent in various complex systems. The authors propose a novel pretopological clustering algorithm that allows for customizable logical clustering and high parametrization, enabling heathcare experts and other professionals to better interpret and utilize the results for clustering, diagnosis and recommendation. PretopoMD performs well on a variety of quality indicators and demonstrates the potential for hierarchical clustering in handling complex data. By providing users with the ability to fine-tune the clustering process using logical rules and parameters, PretopoMD offers a more interpretable and actionable clustering result. However, the paper also acknowledges that the quality of clustering is highly dependent on the user's objectives and the context in which the data is used. This emphasizes the need for a comprehensive understanding of both the dataset and the clustering method, as well as the importance of developing improved visualization tools for high-dimensional contexts.

Future work will focus on exploring these solutions while working on a large and rich dataset comprising healthcare consumption, weather time series, and mixed building characteristic typologies. By collaborating with field experts and analyzing the significance of the clusters, the authors aim to adjust parameters and logical rules accordingly, ensuring the resulting clusters are meaningful and useful for specific case studies. Additionally, investigating automated hyperparameter tuning based on a combination of quality indicators will be pursued to further improve the clustering process for mixed data containing time series.

## REFERENCES


[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. Time-series clustering–a decade review. *Information systems* 53 (2015), 16–38.

[2] Amir Ahmad and Lipika Dey. 2011. A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognition Letters* 32, 7 (2011), 1062–1069.

[3] Amir Ahmad and Shehroz S Khan. 2019. Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access* 7 (2019), 31883–31902.

[4] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When is "nearest neighbor" meaningful?. In *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*. Springer, 217–235.

[5] Christophe Biernacki. 2016. BigStat for Big Data: Big Data clustering through the BigStat SaaS platform. In *Journée scientifique Big Data & Data science*.

[6] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.

[7] David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), 224–227.

[8] William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification* 1, 1 (1984), 7–24.

[9] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment* 1, 2 (2008), 1542–1552.

[10] B Escofier. 1979. Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahiers de l'Analyse des Données* 4, 2 (1979), 137–146.

[11] Alex Foss, Marianthi Markatou, Bonnie Ray, and Aliza Heching. 2016. A semi-parametric method for clustering mixed data. *Machine Learning* 105, 3 (2016), 419–458.

[12] Ben D Fulcher, Max A Little, and Nick S Jones. 2013. Highly comparative time-series analysis: the empirical structure of time series and their methods. *Journal of the Royal Society Interface* 10, 83 (2013), 20130048.

[13] Sami Mohammad Halawani, M Alhaddad, and A Ahmad. 2012. A study of digital mammograms by using clustering algorithms. (2012).

[14] Zhexue Huang. 1997. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. Citeseer, 21–34.

[15] Zhexue Huang. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery* 2, 3 (1998), 283–304.

[16] Adil Mehmood Khan, Young-Koo Lee, Sungyoung Y Lee, and Tae-Seong Kim. 2010. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE transactions on information technology in biomedicine* 14, 5 (2010), 1166–1172.

[17] Julio Laborde. 2019. *Pretopology, a mathematical tool for structuring complex systems: methods, algorithms and applications.* Ph. D. Dissertation. Paris Sciences et Lettres (ComUE).

[18] Elizabeth Ann Maharaj. 2000. Cluster of Time Series. *Journal of Classification* 17, 2 (2000).

[19] Ian C McDowell, Dinesh Manandhar, Christopher M Vockley, Amy K Schmid, Timothy E Reddy, and Barbara E Engelhardt. 2018. Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS computational biology* 14, 1 (2018), e1005896.

[20] Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 33–42.

[21] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[22] Dharmendra S Modha and W Scott Spangler. 2003. Feature weighting in k-means clustering. *Machine learning* 52, 3 (2003), 217–237.

[23] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[24] Tim Oates, Laura Firoiu, and Paul R Cohen. 1999. Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, Vol. 17. Citeseer, 21.

[25] Julio-Omar Palacio-Niño and Fernando Berzal. 2019. Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667* (2019).

[26] Arkanath Pathak and Nikhil R Pal. 2016. Clustering of mixed data by integrating fuzzy, probabilistic, and collaborative clustering framework. *International Journal of Fuzzy Systems* 18 (2016), 339–348.

[27] G Philip and BS Ottaway. 1983. Mixed data cluster analysis: an illustration using Cypriot hooked-tang weapons. *Archaeometry* 25, 2 (1983), 119–133.

[28] Claudia Plant and Christian Böhm. 2011. Inconco: interpretable clustering of numerical and categorical objects. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*. 1127–1135.

[29] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[30] Skyler Seto, Wenyu Zhang, and Yichen Zhou. 2015. Multivariate time series classification using dynamic time warping template selection for human activity recognition. In *2015 IEEE symposium series on computational intelligence*. IEEE, 1399–1406.

[31] Michael Steinbach, Levent Ertöz, and Vipin Kumar. 2004. The challenges of clustering high dimensional data. *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition* (2004), 273–309.

[32] CB Storlie, SM Myers, SK Katusic, AL Weaver, RG Voigt, PE Croarkin, RE Stoeckel, and JD Port. 2018. Clustering and variable selection in the presence of mixed variable types and missing data. *Statistics in medicine* 37, 19 (2018), 2884–2899.

[33] Michel Verleysen and Damien François. 2005. The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired Systems: 8th International Work-Conference on Artificial Neural Networks, IWANN 2005, Vilanova i la Geltrú, Barcelona, Spain, June 8-10, 2005. Proceedings 8*. Springer, 758–770.

[34] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *J. Mach. Learn. Res.* 22, 201 (2021), 1–73.

[35] Fatin N Zainul Abidin and David R Westhead. 2017. Flexible model-based clustering of mixed binary and continuous data: application to genetic regulation and cancer. *Nucleic acids research* 45, 7 (2017), e53–e53.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009