Space Time Series Clustering: Algorithms, Taxonomy, and Case Study on Urban Smart Cities

Asma Belhadi, Youcef Djenouri, Kjetil Nørvåg, Heri Ramampiaro, Florent Masseglia, Jerry Chun-Wei Lin

To cite this version:


HAL Id: lirmm-03036868
https://hal-lirmm.ccsd.cnrs.fr/lirmm-03036868

Submitted on 2 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

This paper provides a short overview of space time series clustering, which can be generally grouped into three main categories such as: hierarchical, partitioning-based, and overlapping clustering. The first hierarchical category is to identify hierarchies in space time series data. The second partitioning-based category focuses on determining disjoint partitions among the space time series data, whereas the third overlapping category explores fuzzy logic to determine the different correlations between the space time series clusters. We also further describe solutions for each category in this paper. Furthermore, we show the applications of these solutions in an urban traffic data captured on two urban smart cities (e.g., Odense in Denmark and Beijing in China). The perspectives on open questions and research challenges are also mentioned and discussed that allow to obtain a better understanding of the intuition, limitations, and benefits for the various space time series clustering methods. This work can thus provide the guidances to practitioners for selecting the most suitable methods for their used cases, domains, and applications.

Keywords: Data mining, Clustering, Space time series
1. Introduction

Recent advances in geolocation, partly as a result of GPS (Global Positioning System) support, has resulted in the creation of large volumes of data varied in time and space. Space time series is one of the most powerful representations for several domains applications including transportation J. Hu et al. 2018, health-care Xi et al. 2018, seismology Morales-Esteban et al. 2014 and climate science Y. Liu 2015. The useful way to analyze space time series is by utilizing data mining and machine learning techniques Izakian, Pedrycz, and Jamal 2013; Von Landesberger et al. 2016. Clustering is one of the data mining techniques where similar data are grouped together into homogeneous clusters that has been intensively studied in the past decades Karypis, E.-H. Han, and Kumar 1999; Ng and J. Han 2002; Vesanto and Alhoniemi 2000; Karypis 2002; Nanopoulos, Theodoridis, and Manolopoulos 2001; H. Xiong, J. Wu, and J. Chen 2009; Ester et al. 1996; Jain, Murty, and Flynn 1999; Kriegel, Kröger, and Zimek 2009. In recent decades, many research focused on time series clustering Keogh and J. Lin 2005; Hallac et al. 2017; Dau, Begum, and Keogh 2016; Y. Xiong and Yeung 2002, and several works considered the spatial dimension in time series clustering Ferstl et al. 2017; Gharavi and B. Hu 2017; Izakian, Pedrycz, and Jamal 2015; Izakian, Pedrycz, and Jamal 2013, resulting in space time series clustering.

This paper presents a comprehensive overview of the existing space time series clustering algorithms. We have divided the existing approaches into three main categories depending on the type of clustering results. The first category is called hierarchical space time series clustering that is used to create hierarchical clusters among the space time series data. The second category is named pure partitioning space time series clustering that is utilized to partition the space time series into disjoint and similar clusters. For the third overlapping partitioning space time series clustering, it aims at determining clusters where space time series data may belong to one or more clusters. In this paper, we then study and present the solutions for each category. In addition, we show the applications of existing space time series clustering on urban traffic data relevant to two smart cities (e.g., Odense in Denmark and Beijing in China). Furthermore, challenges, open perspectives and research trends for space time series clustering are discussed and concluded. Compared to previous survey papers, this paper first provides
a deep analysis of space time series clustering techniques, which allows to clearly understand the merits and the limits of the reviewed algorithms for each space time series clustering category. This paper also derives mature solutions for space time series clustering, in particular for massive data, and for emerging applications.

1.1. Previous studies

This section summarizes the relevant survey papers and clarifies the differences to show the contributions of this paper. This survey paper is composed of two main topics, which are spatio-temporal data mining and time series clustering. In the following section, we review some existing surveys of these topics. Many data mining approaches have been proposed for spatio-temporal data.

Zheng 2015 reviewed trajectory data mining techniques including clustering, classification, and outlier detection. Feng and Zhu 2016 proposed an overall framework of trajectory data mining including preprocessing, data management, query processing, trajectory data mining tasks, and privacy protection. Shekhar, Evans, et al. 2011 and Gupta et al. 2014 provided the comprehensive overviews of application-based scenarios for spatio-temporal data mining such as financial markets, system diagnosis, biological data, and user-action sequences. Eftelioglu et al. 2016 studied hot spot detection in several applications such as environmental criminology, epidemiology, and biology. Keogh and Kasetty 2003 introduced the need of a fair evaluation of time series data including time series clustering. According to the authors, such as evaluation is done to avoid data and implementation bias. Liao 2005 presented an overview of time series clustering. It categorizes time series clustering into three categories, which are i) raw-data-based approaches either in time or frequency domain; ii) feature-based approaches that use feature extraction techniques for handling high dimensional time series reduction; and iii) model-based approaches that each time series is obtained by applying some mixture of models. Zolhavarieh, Aghabozorgi, and Teh 2014 reviewed the existing works of subsequence time series clustering based on the published periods such as: preproof (1997–2003), interproof (2003–2010), and postproof (2011–2014).

Fu 2011 discussed time series data mining techniques including segmentation, indexing, clustering, visualization and pattern discovery. Esling and Agon 2012 reviewed the existing time series data mining approaches such as classification, clustering, segmentation, outlier detection, prediction, and
rules and motifs discovery. It classifies time series clustering into two categories, which are i) whole series clustering by considering the complete time series in the clustering process, and ii) sub-sequences clustering, in which the clusters are found by selecting subsequences from multiple time series. In addition, Aghabozorgi, Shirkhorsidi, and Wah 2015 included another category of time series clustering, namely time point clustering, which aims at determining clusters based on a combination of the temporal proximity of time points and the similarity of the corresponding values. Compared to the existing surveys, this is the first survey that deals with space time series data; all the other works have been limited to only time series data, or even to spatial or temporal data.

<table>
<thead>
<tr>
<th>Class</th>
<th>Algorithms</th>
<th>Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical</td>
<td>Agglomerative Clustering</td>
<td>Rodriguez and Laio 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shen and Cheng 2016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X. Wang et al. 2019</td>
</tr>
<tr>
<td></td>
<td>Hierarchical Self Organizing Map</td>
<td>Steiger, Resch, and Zipf 2016</td>
</tr>
<tr>
<td></td>
<td>Machine Learning</td>
<td>Y. Wu et al. 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ferstl et al. 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Deng et al. 2018</td>
</tr>
<tr>
<td>Pure Partitioning</td>
<td>kmeans</td>
<td>N. Andrienko and G. Andrienko 2013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gharavi and B. Hu 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cho et al. 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H.-L. Yu et al. 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X. Jiang, C. Li, and J. Sun 2018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bai et al. 2014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Krüger et al. 2017</td>
</tr>
<tr>
<td></td>
<td>PAM</td>
<td>Von Landesberger et al. 2016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Penfold et al. 2016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T. Sun et al. 2017</td>
</tr>
<tr>
<td></td>
<td>Peak Density</td>
<td>J. Jiang, Y. Chen, et al. 2019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Putri et al. 2019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Heredia and Mor 2019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H. Li 2019</td>
</tr>
<tr>
<td>Overlapping Partitioning</td>
<td>Fuzzy kmeans</td>
<td>Izakian, Pedrycz, and Jamal 2013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Izakian, Pedrycz, and Jamal 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disegna, D’Urso, and Durante 2017</td>
</tr>
<tr>
<td></td>
<td>Machine learning</td>
<td>Paci and Pinazzi 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gholami and Pavlovic 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y. Zhang et al. 2017</td>
</tr>
</tbody>
</table>

Table 1: Taxonomy of space time series clustering algorithms

1.2. Taxonomy and paper organization

Table 1 presents a taxonomy of the space time series clustering algorithms presented in this paper. They are classified into three categories. The first
category is named *hierarchical clustering*, which is utilized to identify hierarchy among the space time series data. The second *pure partitioning space time series clustering* category is to partition the space time series into disjoint and similar clusters. Furthermore, the third *overlapping partitioning space time series clustering* category aims at determining a space time series that may belong to one or more clusters.

The rest of the paper is organized as follows. Section 2 defines the background and concepts used in the paper, including clustering and space time series data. Section 3 presents the relevant approaches for space time series algorithms. Section 4 shows a case study of the existing space time series clustering algorithms on large and big urban traffic data by exploring two urban smart cities (Odense in Denmark and Beijing in China). Section 5 discusses the challenges and future directions in space time series clustering. Finally, Section 6 states the conclusion of this paper.

2. Preliminaries

This section presents preliminaries regarding clustering techniques and space time series data.

2.1. Clustering

**Definition 1 (Clustering).** Consider $m$ data $x_1, x_2, \ldots, x_m$, and a set of $k$ clusters $C = \{C_1, C_2, \ldots, C_k\}$, and a distance measure $D$. Each cluster $C_i$ is represented by its centroid $g_i$. Any clustering algorithm aims to partition the data into similar groups such as the optimal clustering denoted as $C^*$:

$$C^* = \arg\min_{C} \sum_{i=1}^{k} \sum_{x_j \in C_i} D(g_i, x_j)$$  \hspace{1cm} (1)

**Definition 2 (Hierarchical Clustering).** Hierarchical clustering aims to create a tree-like nested structure partition $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_h\}$ of the data such that:

$$\forall (i, j) \in [1 \ldots k]^2, \forall (m, l) \in [1 \ldots h]^2, C_i \in \mathcal{H}_m, C_j \in \mathcal{H}_l, m \geq l \Rightarrow C_i \in C_j \land C_i \cap C_j = \emptyset$$

A hierarchical algorithm builds the hierarchical relationship among data. The typical approach is that each data point is first in an individual cluster. Based on the most neighboring, the clusters are merged to new clusters.

Definition 3 (Pure Partitioning Clustering). Pure partitioning clustering aims to look for $k$ partitions of the data such that:

\[ C_i \cap C_j = \emptyset, \text{ and } \forall i \in [1\ldots k], C_i \neq \emptyset \]

The basic idea of pure partitioning clustering is to consider the center of data points as the center of the corresponding cluster, and recursively compute and update the center until convergence criterion is achieved. Typical algorithms of this kind of clustering include $k$-means MacQueen et al. 1967, PAM (Partition Around Medoids) Kaufman and Rousseeuw 1990, and CLARA (Custering LARge Applications) Kaufman and Rousseeuw 2009.

Definition 4 (Overlapping Partitioning Clustering). Overlapping partitioning clustering indicates that each data $x_j$ to each cluster $C_i$ is with a degree of membership $\mu_{ij} \in [0\ldots 1]$ such that $\sum_{i=1}^{k} \mu_{ij} = 1$

The basic idea of overlapping clustering is to assign data point to each cluster using a membership value between 0 and 1 in order to describe the relationship between data points and clusters. Typical algorithms of this kind of clustering include fuzzy $c$-means Bezdek, Ehrlich, and Full 1984, and FCS (Fragment Clustering Schemes) Dave and Bhaswan 1992.

2.2. Space Time Series

Definition 5 (Space Time Series). Consider $m$ data such that $x_1, x_2, \ldots, x_m$, each of data is comprised of a spatial part and a time series part. For the $l^{th}$ data $x_l$, the concatenation $x_l = [x_l(s)|x_l(t)]$ is realized, where $x_l(s)$ represents its spatial part and $x_l(t)$ refers to its time series part. By considering $r$ features (usually, $r = 2$) for the spatial part and $q$ features for the time series part, we have the following representation for the $l^{th}$ data with dimensionality $n = r + q$

\[ x_l = [x_{l1}(s), \ldots, x_{lr}(s)|x_{l1}(t), \ldots, x_{lq}(t)] \] (2)
Figure 1 illustrates an example of a space-time series. There are a number of spatial points in \( x-y \) coordinates, and for each spatial point, there is one (or more) time series representing the measurements of a phenomenon in different time steps.

**Definition 6 (Space Time Series Similarity).** We define the distance between two space time series \( x_1 \) and \( x_2 \) as:

\[
D(x_1, x_2) = SD(x_1, x_2) + TD(x_1, x_2),
\]

where i) \( SD(x_1, x_2) \) defines the spatial distance that computes the similarity between the spatial components of the space time series \( x_1 \), and \( x_2 \), and ii) \( TD(x_1, x_2) \) defines the temporal distance that determines the similarity between the time series components of \( x_1 \), and \( x_2 \). The spatial distance is usually computed using ordinary Euclidean distance, where the temporal distance is captured using time series distances Mori, Mendiburu, and Lozano 2016. In the following, we illustrate some interesting measures between two time series data (e.g., \( x(t) \) and \( y(t) \)).

1. \( L_p \) distances Yi and Faloutsos 2000: \( L_p \) distances are the rigid metrics that can only compare series of the same length. However, due to their
simplicity, they have been widely used in many tasks related to time series analysis and mining. The different variations of the $L_p$ distances and their formulas are provided in Table 2.

$$
\begin{align*}
\text{Table 2: } L_p \text{ Distances} \\
\text{Distance} & \quad p & \quad \text{Formula} \\
\text{Manhattan} & \quad 1 & \quad \sum_{x_i(t) \leq x(t) \lor y_i(t) \geq y(t)} |x_i(t) - y_i(t)| \\
\text{Euclidean} & \quad 2 & \quad \sqrt{\sum_{x_i(t) \leq x(t) \lor y_i(t) \geq y(t)} (x_i(t) - y_i(t))^2} \\
\text{Minkowski} & \quad [1\ldots\infty] & \quad \max_{x_i(t) \leq x(t) \lor y_i(t) \geq y(t)} \{ |x_i(t) - y_i(t)| \} \\
\text{Infinite norm} & \quad \infty & \quad \max_{x_i(t) \leq x(t) \lor y_i(t) \geq y(t)} \{ |x_i(t) - y_i(t)| \}
\end{align*}
$$

Table 3: Time Series Distances

$$
\begin{align*}
\text{Distance} & \quad \text{Formula} \\
\text{DTW}(x(t), y(t)) & \quad \begin{cases} 
0 & \text{if } |x(t)| - 1 = |y(t)| - 1 = 0 \\
\infty & \text{if } |x(t)| - 1 = 0 \text{ or } |y(t)| - 1 = 0 \\
x_0(t) - y_0(t) + \min\{\text{DTW}(x(t)/x_0(t), y(t)/y_0(t)), \text{DTW}(x(t), y(t)/y_0(t))\} & \text{otherwise}
\end{cases} \\
\text{STS}(x(t), y(t)) & \quad \sqrt{\sum_{x_i(t) \leq x(t) \lor y_i(t) \geq y(t)} \left( \frac{y_{i+1}(t) - y_i(t)}{t_{i+1} - t_i} - \frac{x_{i+1}(t) - x_i(t)}{t_{i+1} - t_i} \right)^2} \\
\text{Dissim}(x(t), y(t)) & \quad \sum_{x_i(t) \leq x(t) \lor y_i(t) \geq y(t)} (x_i(t) - y_i(t) + x_{i+1}(t) - y_{i+1}(t)) \times (t_{i+1} - t_i) \\
\text{PC}(x(t), y(t)) & \quad \frac{\sum_{x_i(t) \leq x(t) \lor y_i(t) \geq y(t)} (x_i(t) - x(t)) \times (y_i(t) - y(t))}{\sqrt{(x_i(t) - x(t))^2} \times \sqrt{(y_i(t) - y(t))^2}}
\end{align*}
$$

Note that $\bar{x}(t)$ and $\bar{y}(t)$ are the mean values of time series $x(t)$ and $y(t)$, respectively.

2. DTW (Dynamic Time Warping) Sakoe and Chiba 1978: This distance is able to deal with transformations such as local warping and shifting. Furthermore, it allows the comparison between different series length.
3. STS (Short Time Series) Möller-Levet et al. 2003: This distance is adapted to the characteristics of irregularly sampled series.
4. Dissim Frentzos, Gratsias, and Theodoridis 2007: It is specifically designed for series collected at different sampling rates that indicates each series is defined in a finite set of time instants, but these can be different for each series.
5. PC (Pearson’s correlation) Golay et al. 1998: This distance focuses on extracting a set of features from the time series and calculating the similarity between these features instead of using the raw values of the series.

Table 3 presents the description of DTW, STS, Dissim, and PC formulas. The interested readers may also find the alternative distances with
SAX (Symbolic Aggregate approXimation) Cole, Shasha, and Zhao 2005 or sketches Shieh and Keogh 2008 but until now, we do not find those relevant works in space time series clustering.

3. Algorithms

Intensive studies have been carried to capture the space time series clustering algorithms. Up to now, 112 papers have been analyzed. From these papers, the following filter process is performed:

1. 28 papers are removed after the first pre-screening of the abstract due to they are out of the scope of space time series clustering.
2. From the remaining 84 papers, only 32 papers are finally selected. The selection criteria is based on the quality of the paper, and the quality of the publisher. Only high quality papers are elected if the paper is published in top-tier conferences or high impact factor journals.

In addition, our intensive study reveals that almost solutions for space time series clustering can be classified as hierarchical-based, pure partitioning-based, and overlapping-based approaches. Therefore, this section presents space time series clustering algorithms grouped into three categories.

3.1. Space-time series hierarchical clustering

**Before 2017:** Rodrigues, Gama, and Pedroso 2008 presented ODAC (Online Divisive Agglomerative Clustering) incrementally updates the tree-like hierarchy clusters using top-down strategy by computing the dissimilarities between time series. The dissimilarities between time series is defined by the cluster’s diameter measure, where the splitting criterion is supported by a confidence level given by the Hoeffding bound Hoeffding 1963 that allows to define the diameter of each cluster. This algorithm is applied to multiple time series, but it can be easily adopted on space time series data. Shen and Cheng 2016 established a new framework that enables comprehensive analysis of trajectory space-time series data to group people with similar behavior. The framework segregates individuals into subgroups upon where (place), when (time) and how long (duration) the activities are conducted for each individual. It includes three main steps, (i) extracting ST-ROI, i.e., the region of interests with not only spatial location information but also interesting time spans; (ii) defining individual time allocation on the ST ROIs as their space-time profiles to describe his/her activity routine; and (iii) group
people using hierarchical clustering based on different activity patterns. For example, Steiger, Resch, and Zipf 2016 develop the geographic hierarchical self-organizing map (Geo-H-SOM) for discovering spatio-temporal semantic clusters from tweets of people generated in different week days and on different regions. It considers the similarities between tweets across their temporal as well as geographical and semantic characteristics. Results on real case study in London reveal similar correlations between tweets of people live in the same region. However, some clusters are sparse and difficult to be analyzed due to high correlation between certain tweets.

From 2017: Ferstl et al. 2017 propose a hierarchical ensemble clustering approach to analyze and visualize temporal uncertainty in weather forecasting data. Clusters in specified time window are merged to indicate when and where forecast trajectories diverge. Different visualizations of time-hierarchical grouping on European center for medium-range weather forecasting data are shown including space-time surfaces built by connecting cluster representatives over time, and stacked contour variability plots. Y. Wu et al. 2017 presented an interactive framework named StreamExplorer that visualizes social streams. It continuously detects important time periods (i.e., sub-events), and extract topics of tweets made on any sub-events using GPU-assisted self organizing map. A multi-level visualization method that integrates Agnes algorithm for showing a space-time series generated from the extracted tweets in a given time period and for different users located on different regions. The map allows to summarize important sub-events at a macroscopic level using a tree of visualizations. It not only reveals the dynamic changes of a social stream in the context of its past evolution, but also organizes historical sub-events in a hierarchical manner for easy review and navigation of sub-events. The proposed system enables end users to track, explore, and gain insights of social streams at different levels. Deng et al. 2018 address a spatio-temporal heterogeneity problem by employing space-time series clustering. This approach divides space-time series data into meaningful clusters while considering both the spatial proximity and the time series similarity instead of the previous methods that only deal with time or space dimensions. The application of auto-correlation time-series clusters in artificial neural networks reveals good accuracy for space-time series prediction. X. Wang et al. 2019 suggested a novel representation formed by a sequence of 3-tuples for interval-valued time series in high dimensional data, and loss information issue. In addition, a hierarchical clustering algorithm based on improved dynamic time warping distance measure is designed for
interval-valued time series of equal or unequal length.

3.2. Space-time series pure partitioning clustering

**Before 2016:** N. Andrienko and G. Andrienko 2013 proposed an interactive framework to analyze and visualize large amount of spatio-temporal data represented by a set of time series. Multiple and heterogeneous space-time series are first created from the spatio temporal data. The set of space-time series is then grouped based on the similarity of the temporal variation of the timestamp value. During this step, an interactive tool is used to refine the clustering results. This is done by showing both time graph and map displays to the data analysts. The time-graph display let the analysts view the homogeneity degree of each group. For the map display, the locations are characterized by the space-time series and each group is painted in the same color. Cho et al. 2014 develop the Stroscope visualization tool that help neurologists analyze space-time series coming from blood pressure data. It provides two kinds of clustering techniques such as: data-space clustering and image-space clustering. For the data-space clustering, records that have similar measurement values are grouped together, which result in the same clusters for the same data set. However, an image-space clustering aims at solving the visual inconsistency problem of the data-space clustering. In the image space clustering, records with a similar color pattern are clustered together, where the clustering results could vary according to the color table defined by the users, but the results are more reasonable to users who expresses his/her intention in his/her color mapping choice. Bai et al. 2014 proposed a Gtem algorithm to cluster events from geographical temperature sequence data. It can detect high temperature events in irregular shape, size and evolution model. Furthermore, Gtem can automatically select the optimal parameters based on the MDL (Minimum Length Description) principle Barron, Rissanen, and B. Yu 1998 to automatically group events of an area into space-time series data. Moreover, Gtem can successfully find high-temperature events with exact start-end timestamps on the daily weather of the Hunan province in China from 2004–2008. H.-L. Yu et al. 2015 applied the clustering process to distinguish the space-time patterns of local precipitations in the summer and autumn synoptic conditions from 24 gauges during 1996–2008 in Taiwan. It groups the synoptic and local conditions for the space-time rainfall patterns by integrating k-means with the empirical orthogonal function analysis. The results identified three mainly extreme patterns and two normal patterns in both seasons. L. Li et al. 2015 in-
vestigated trend modeling for space time series clustering in the context of urban traffic data. Based on daily similarity of traffic time series on different urban locations, the simple average trend with PCA (Principle Component Analysis) approach is developed to analyze the daily traffic space time series obtained in consecutive days and define their global varying similarity while DWT (Discrete wavelet transform) mostly defines the local varying tendency.

2016: Von Landesberger et al. 2016 presented a visual analysis for people flow between places in London. The people flow is aggregated into regions to reduce the mass mobility patterns using $k$-means algorithm. Despite of visualization of people flow, only aggregated regions are shown to the users for better understanding the distribution of flow between places. Moreover, a new measure named Strength Flow is developed to filter the regions having low density flow. L. Wang et al. 2016 developed a clustering method called separation degree algorithm that is able to construct self-adaptive interval based on the separation degree model to detect anomaly in network space data. The advantage of this approach is to automatically determine the self-adaptive interval, which can be used to improve the accuracy of anomaly detection. Extensive experimental results showed that the proposed method can effectively detect anomaly data from heterogeneous spaces in the given network. Penfold et al. 2016 suggested a clustering model to identify clusters of early adoption for a new clinical practice. The results indicated that the revealed patterns provide insights to identify organizational context and prescribe level factors involved in diffusion and implementation within a learning health care system. The proposed approach can be used for real-time prospective surveillance context such as urgent clinical events with public health importance. Steiger, Resch, Albuquerque, et al. 2016 used a geographic self-organizing map to group human mobility patterns by analyzing similar space-time series generated from live traffic feeds. A standard self-organizing map is first applied in order to observe and analyze the general topological relationships of the reference database. A geographic self-organizing map is then computed for the identification of similar overlapping traffic disruption patterns. The results of traffic disruption clusters are finally correlated with the computed geo-referenced weight vectors from all retrieved geo-referenced traffic data. A case study in London traffic data showed that particularly special events, such as concerts, demonstrations, and sports events, etc., are well reflected within space-time series input data.

2017: T. Sun et al. 2017 proposed matching and pruning strategies to efficiently compute the center of space-time series using dynamic time warping
distance. Experimental results revealed that the proposed centroid formula improves the performance compared to the existing space-time series clustering in terms of computational time and clustering quality. Gharavi and B. Hu 2017 presented a clustering algorithm to detect disturbances and degradation area in the grid. The $k$-means algorithm is extended to group measurement units into different clusters based on power quality. A multi-objective criterion is defined by considering both time and space in the clustering process. According to the experiments on IEEE 39-bus transmission system, it revealed that the proposed clustering space-time synchrophasor scheme is capable to detect and isolate areas in the grid suffering from multiple disturbances, such as faults. Krüger et al. 2017 proposed a segmentation approach that allows distinct activities within human motion space-time series data. Segmentation human motion data is first considered as graph problem, and a neighborhood graph-based and PCA (Principle Components Analysis) approaches are then applied for dimension reduction. A clustering method that allows to detect motion segments is based on self-similarities which needs no assumption on the number of clusters. The experiments on a wide variety of motion datasets show that the approach can identify usual non-repetitive human activities such as one step, jump, and, turn. However, the approach could not identify some substantial changes between the individual repetitions in the muscle activation patterns such as fatigue effects in longer motion trials.

2018: X. Jiang, C. Li, and J. Sun 2018 focused on mining of multimedia time series data using a mixed composition of graphic arts and pictures, hyper text, text data, video or audio. It adopted a $k$-means algorithm to handle high dimensional data as the input set for a multimedia database and at the same time, the algorithms obtains optimal similarity measure by utilizing a Minkowski distance which is a generalized form of the Euclidean distance. Mikalsen et al. 2018 proposed a robust time series cluster kernel by taking the missing time series data into account using the properties of Gaussian mixture models augmented with informative prior distributions. An ensemble learning approach is exploited to ensure robustness of parameters by combining the clustering results of many Gaussian mixture models to form the final kernel.

2019: Some algorithms based on density peaks principle Rodriguez and Laio 2014 and its variants gravitation-based density peaks clustering J. Jiang, Hao, et al. 2018, and density peaks clustering based on logistic distribution and gravitation J. Jiang, Y. Chen, et al. 2019 have been suggested for space
time series clustering. The main idea behind these algorithms is that the centers of the different clusters are more dense than the remaining data. This allows to automatically identify outliers. In addition, the clusters are recognized regardless of their shape and of the dimensionality of the space in which they are embedded. Y. Wang et al. 2016 implements a time-based Markov model to formulate the dynamics of electricity consumption for customer behaviors by considering the state-dependent characteristics. It also indicates that future consumption behaviors would be related to the current states. Furthermore, it mentions that the density peak clustering has good robustness to identify outliers without further processing. Putri et al. 2019 developed a new density-based clustering approach for grouping a set of time series. This approach generates arbitrarily shaped clusters, and explicitly tracks their temporal evolution. Heredia and Mor 2019 proposed a hybrid approach which combines the density-peak clustering with the spatial density of space time series data. The whole data is first partitioned using the smoothed density function, and the resulted groups are further divided using the density-peak clustering approach. H. Li 2019 developed a hybrid approach based on principal component analysis and density-peak clustering. A high dimensional multi-variate space time series data is first reduced using the principal component analysis, and the selected features are then grouped using the density-peak clustering.

3.3. Space-time series overlapping partitioning clustering

**Before 2017:** Izakian, Pedrycz, and Jamal 2013 proposed a fuzzy approach for spatio-temporal data clustering. Fuzzy c-means and adaptive Euclidean distance function are adopted to cluster different nature of spatio-temporal data. The suggested augmented distance allows to control the effect of each data in the determination of the overall Euclidean distance and gives a sound balance between the impact of the spatial and temporal components of the data. Izakian and Pedrycz 2014b suggested a cluster-center approach for anomaly detection problem in space-time series data. A Fuzzy C-Means (FCM) algorithm is employed to group the time series. A Euclidean distance is used for similarity computation in both spatial and temporal components, where the \( \lambda \) parameter is defined to balance the impact of the spatial and temporal components of the data in the clustering process. In addition, Anomaly score is assigned to each cluster for quantifying the unexpected changes in the structure of data. At the end, the relations between clusters presented in successive time windows are visualized to quantify anomaly propagation over
time. Izakian and Pedrycz 2014a introduced a generalized version of fuzzy c-means clustering to cluster data with blocks of features coming from distinct sources. A new distance function is developed to take the multi-sources aspect into account. The distance combines the features of different sources by using aggregate variables, that allows to increase/decrease the impact of the given data source against other data sources. Izakian, Pedrycz, and Jamal 2015 further presented three alternative approaches for fuzzy clustering of space-time series data. The first approach takes the averaging dynamic time warping distance technique into account and applies the fuzzy c-means technique for clustering space-time series data. For the second approach, a fuzzy c-medoids technique that ignores the average distance calculation was explored and finally, a combination between the c-medoids and the c-medoids was examined and discussed.

From 2017: Paci and Finazzi 2017 developed a Bayesian dynamic approach that integrates a finite weighted mixture model for clustering space-time series data. Thus, a state-space model has been employed to describe the temporal evolution of different locations belonging to each cluster. Also, a new strategy for selecting the number of clusters has presented. By using a weighted mixture model, this approach allows easy and fast prediction of the membership probability at any location and at any window time. Diegna, D’Urso, and Durante 2017 proposed COFUST (COpula-based FUzzy clustering algorithm for Spatial Time series). A combination of Fuzzy Partitioning Around Medoids (FPAM) algorithm Kaufman and Rousseeuw 1990 with a copula-based approach Di Lascio, Durante, and Pappada 2017 is performed to interpret co-movements of large-scale time series. First, both spatial and temporal dependencies between the time series are identified through a copula-based approach. Then, the FPAM algorithm has been adopted in order to determine non-fictitious patterns in the space-time series and producing the final clusters. This approach is computationally more efficient and tend to be less affected by both local optima and convergence problems compared to the existing space-time series overlapping clustering. Gholami and Pavlovic 2017 considered the temporal dependency between space-time series of complex human motion data. The temporal dependencies are modeled using Gaussian process whose covariance function controls the desired dependence. The Bernoulli process is also incorporated into the overall process to concurrently learn the dimensionality of the subspaces from the data. Y. Zhang et al. 2017 introduced a density-contour based spatio-temporal clustering approach (ST-DPOLY) and compare it with the spatio-temporal
shared nearest neighbors (ST-SNN). First, a spatial density function is determined for the spatial point data collected in batches, where a density threshold is used for each batch of time to identify spatial clusters. Spatio-temporal clusters are then determined as continuing clusters. Continuing clusters are defined as the clusters highly correlated in consecutive batches of time. The proposed approach has applied to 1.1 billion taxi trips recorded over seven consecutive years from 2009 to 2016, and presented advantages in terms of clustering results, time and space complexity, while ST-SNN is more interesting in terms of temporal flexibility.

3.4. Discussions

From the above literature review, we provide our insights of the reviewed papers.

1. Clustering of space time series data requires a lot of efforts, especially in terms of a suitable treatment of the spatial and temporal components of the data. Existing space time series clustering algorithms have been developed in this direction. Still, much further work is needed to achieve mature solutions. For instance, current algorithms consider temporal and spatial dimensions in the same processing level. However, in some cases, temporal dimension is more suitable than the spatial one, and vice versa. One way to tackle this issue is to transform the space time series clustering as multi-objective optimization problem, where some aggregation functions may be used between the different dimensions (spatial and temporal dimensions in this case).

2. Space time series data are usually gathered from sensors, and should be processed continually in a data streams environment. The major concern with the existing clustering time series is that they do not provide mechanism to deal with data streams. Incremental clustering algorithms may be an alternative solution. The main merit of these algorithms is that it is not necessary to store the whole data in the memory. Thus, the space requirement of incremental algorithms is relevant small. Typically, they are non-iterative; their time requirement is also small. Adopting incremental clustering algorithms in a space time series is beneficial to practitioners for dealing with more real-world applications relevant to manufacturing and smart city, among others.
In addition, we present the merits and limitations of the existing space time series algorithms (See Table 4 for more details). We can classify the space time series algorithms into three groups, according to various clustering models:

1. Algorithms in the first group aim at finding hierarchical clusters. They do not need any parameter as the input (i.e., the number of clusters). In addition, it is possible to examine partitions at different granularity levels. However, with large scale data, they require higher computation and a huge memory usage. However, space time series data is normally large scale, thus this model is not suitable for real-world situations.

2. The purpose of the algorithms in second group is to find the disjoint partitions. These algorithms are fast compared to the algorithms in the first category. They are thus more suitable for large scale space time series data. Nevertheless, those algorithms require parameter setting (i.e., the number of clusters), which is normally hard to decide, in particular while considering more dimensions in space time series data.

3. Algorithms in the third group are overlapping partitioning algorithms, which are used to find the overlapping partitions by using a membership degree as input. These algorithms are slow compared to pure partitioning algorithms due to the complexity of the space time series data. Moreover, they are very sensitive to the membership rate and the number of clusters. In addition, to determine the overlapping clusters, they do not distinguish the spatial and temporal dimensions, which degrades the accuracy performance. In many real-world cases, some data may overlap in one dimension but are disjoint in the other. In such cases, overlapping partitioning algorithms would fail to determine the optimal clustering.

<table>
<thead>
<tr>
<th>Category</th>
<th>Merits</th>
<th>Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical Clustering</td>
<td>Free-parameters</td>
<td>High time and memory consuming</td>
</tr>
<tr>
<td></td>
<td>Different level of granularities</td>
<td></td>
</tr>
<tr>
<td>Pure Clustering</td>
<td>Fast time and memory consuming</td>
<td>Difficult to fix the number of clusters</td>
</tr>
<tr>
<td>Overlapping Clustering</td>
<td>Finding Overlapping clusters</td>
<td>High time consuming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Need parameters adjustments</td>
</tr>
</tbody>
</table>
4. Evaluation

In this section, a performance evaluation of space time series clustering is provided. Both standard time series databases \(^1\) and a real case study on urban traffic intelligent transportation are analyzed as follows.

4.1. Case Study: Urban traffic Intelligent Transportation

With the popularization of GPS and IT devices, urban traffic flow analysis has attracted growing attention in the last decades. Zheng 2015 and Feng and Zhu 2016 reviewed spatio-temporal urban data mining techniques. The surveys included segmentation and clustering, detecting outliers and anomaly flows, classification sub-trajectories, and finding frequent and periodical sequential patterns from clusters of trajectories. The traffic flow is computed by counting the number of objects (i.e., cars, passengers, taxis, buses, etc.) across a given location during a time interval. This generates a high number of time series captured in different locations of the urban city. A trivial way to represent these time series captured in different locations is space time series data. Space time series data mining is largely used in many number of domains related to intelligent transportation Jensen et al. 2016; Feng and Zhu 2016. They are used to adapt classical data mining techniques and propose new methods for discovering useful knowledge from urban traffic space time series data. Recent research works of space time series data mining techniques for urban traffic data including clustering, pattern mining, and outlier detection can be found in Shekhar, Evans, et al. 2011; Zhou, Shekhar, and Ali 2014; Koperski, Adhikary, and J. Han 1996; Gupta et al. 2014; Shekhar, Z. Jiang, et al. 2015; Y. Djenouri, Belhadi, J. C.-W. Lin, D. Djenouri, et al. 2019; Y. Djenouri, Belhadi, J. C.-W. Lin, and Cano 2019; Y. Djenouri and Zimek 2018; Y. Djenouri and Zimek 2018. One application of space time series data mining for urban data is clustering. The goal is to find out the similar clusters of urban traffic flows represented in different locations. This section shows a case study of an application of space time series clustering algorithms for dealing with urban traffic data. In the experiments, we consider various clustering algorithms with different similarity measures on two urban traffic data (large dataset for Odense city in Denmark and big dataset for Beijing city in China).

\(^1\)https://archive.ics.uci.edu/ml/index.php
4.2. Datasets

Two real Odense and Beijing traffic flow data have been used for evaluation. These datasets are varied in terms of the number of flow values. The Odense traffic data is considered as a large dataset, where the Beijing traffic data is considered as big dataset. The detailed explanation of these two datasets is given as follows:

**Odense Traffic Data:** The first data is captured from several test locations throughout the Odense city. Each data entry contains information related to the vehicle or bike detected at specific locations such as: gap, length, date, time, speed, and class (i.e., type of vehicle or bike). The location is represented by latitude and longitude dimensions. The speed is calculated by km/h, and the datetime represents the year, the month, the day, the hour, the minute and the second that the vehicle or bike is passed by the given location. The most important information of each vehicle or bike is given as follows:

- **datetime:** It represents the time that the vehicle or bike passed on the location, and the format is: YYYY-MM-DD hh:mm:ss.
- **speed:** It defines the actual speed of the vehicle or bike where it is across the location.
- **class:** It defines the type of vehicle or bike. For example, if the class is set as 2, it represents a passenger car.

For ten locations, sensor infrastructure has been installed in a pilot experiment. The ten locations have different characteristics (i.e., traffic density, counters for cars/bikes) as described in Table 5. The traffic data were obtained between January 1\(^{st}\), 2017 and 30\(^{th}\) April 2018. It consists of more than 12 million vehicles and bikes. The data is made by Odense Kommune \(^2\), and may be retrieved at [http://dss.sdu.dk/projects/its/fpd-lof.html](http://dss.sdu.dk/projects/its/fpd-lof.html).

**Beijing Traffic Data:** The second one is a real urban traffic data obtained from Beijing traffic flow, and retrieved from \(^3\). It consists of more than 3 billion traffic flow values during a two-months time period on more

\(^2\)https://www.odense.dk/
\(^3\)https://www.beijingcitylab.com/
than one hundred locations. The most important information of each car is given as follows:

- **datetime**: It represents the time that the car passed on the location, and format is: YYYY-MM-DD hh:mm:ss.

- **Class**: It defines the type of vehicle or bus.

Figure 2 presents the distribution of urban traffic data among Odense and Beijing cities.

### 4.3. Tool

In this work, we used the algorithms that implemented in SPMF library \(^4\) to perform the space time series clustering. The \(k\)-means, DBSCAN, fuzzy \(c\)-means, CHA, and Density-Peak are the well-known clustering algorithms, and cover all the categories of space time series clustering. Therefore, these algorithms are chosen for performance evaluation. This library is first proposed for pattern mining discovery Fournier-Viger et al. 2014 that has been extended to different data mining applications. It also provides algorithms for analyzing time series data such as SAX J. Lin, Keogh, Wei, et al. 2007 and PAA J. Lin, Keogh, Lonardi, et al. 2003, among others. SPMF is an open-source library implemented in Java. The current version of this tool is v2.42b and was released on 11-th, March, 2020. It currently contains 196 data mining algorithms. We then used SPMF for space time series clustering algorithms by implementing the distances related to space time series data, adding class to space time series representation, and adapting the existing

---

4http://www.philippe-fournier-viger.com/spmf/
clustering algorithms provided in SPMF library for space time series representation. We evaluate algorithms regarding three different dimensions as follows:

1. **Runtime performance**: We perform the runtime of each space time series clustering algorithm including preprocessing step, computing similarity, and determining clusters.

2. **Quality of clusters**: We evaluate the quality of the clusters by two ways. The first way aims to compute the Error Sum of Squares (ESS).
It is the sum of the squared differences between each space time series data and the mean of its group. It can be used as a measure for variation within a cluster. If all cases within a cluster are identical, the ESS is equal to 0. A better clustering result obtains lower ESS value. The ESS formula is given as:

$$ESS = \sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j - \overline{C_i})^2,$$

where $\overline{C_i}$ is the mean value of the space time series data belonging to the cluster $C_i$.

The second way aims to evaluate the quality of clusters for the classification of traffic flow Sumit and Akhter 2019; Rezaei and X. Liu 2019; Qu et al. 2019. The data labels are created for each time series data based on the daily observed traffic. We have obtained three different labels (WD: data for weekday, ST: data for Saturday, and SN: data for Sunday). We have created two files; the first file contains the data without labels, and the second file contains data with labels. We apply space time series clustering techniques on the first file and set the number of clusters as 3 (for DBSCAN and CHA algorithms). We have adjusted their parameters to find 3 clusters in order to make a fair comparison with $k$-means, and fuzzy c-means algorithms. After construction of clusters, we compare each cluster for the data with the same label, and we compute the number of corrected classified data for each cluster as the maximum number of common data between this cluster and each label. We then computed the classification ratio of each algorithm to see evaluate the performance.

For both ways, we used the $k$-fold cross-validation technique, which is largely used in the machine learning community Anand, Kirar, and Burse 2013. This approach involves randomly dividing the set of observations into $k$ groups or folds of approximately equal size. The first fold is treated as a validation set and the model is fit to the remaining folds. The procedure is then repeated $k$ times, where a different group is treated as the validation set.

3. **Memory usage**: We compute the memory consumption of the space time series clustering algorithms by using the *MemoryLogger* provided in SPMF tool.
Table 6: Comparison of the space time series clustering in terms of runtime (seconds), the quality of returned clusters (ESS), and the memory usage (MB) using standard time series databases.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset</th>
<th>CPU</th>
<th>ESS</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>kmeans</td>
<td>Air Quality</td>
<td>25</td>
<td>1.80</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Appliances energy prediction</td>
<td>29</td>
<td>2.21</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>EEG Eye State</td>
<td>31</td>
<td>2.25</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Real-time Election</td>
<td>35</td>
<td>2.51</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Beijing Multi-Site Air-Quality</td>
<td>38</td>
<td>2.59</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Beijing PM2.5</td>
<td>42</td>
<td>2.71</td>
<td>45</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>Air Quality</td>
<td>27</td>
<td>1.71</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Appliances energy prediction</td>
<td>31</td>
<td>2.25</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>EEG Eye State</td>
<td>32</td>
<td>2.28</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Real-time Election</td>
<td>37</td>
<td>2.35</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Beijing Multi-Site Air-Quality</td>
<td>39</td>
<td>2.37</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Beijing PM2.5</td>
<td>41</td>
<td>2.95</td>
<td>47</td>
</tr>
<tr>
<td>Fuzzy cmeans</td>
<td>Air Quality</td>
<td>25</td>
<td>1.80</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Appliances energy prediction</td>
<td>33</td>
<td>2.57</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>EEG Eye State</td>
<td>41</td>
<td>2.29</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Real-time Election</td>
<td>45</td>
<td>2.99</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Beijing Multi-Site Air-Quality</td>
<td>49</td>
<td>3.11</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Beijing PM2.5</td>
<td>55</td>
<td>3.05</td>
<td>52</td>
</tr>
<tr>
<td>CHA</td>
<td>Air Quality</td>
<td>42</td>
<td>0.99</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Appliances energy prediction</td>
<td>43</td>
<td>1.05</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>EEG Eye State</td>
<td>49</td>
<td>1.12</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Real-time Election</td>
<td>56</td>
<td>1.19</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Beijing Multi-Site Air-Quality</td>
<td>63</td>
<td>1.68</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Beijing PM2.5</td>
<td>67</td>
<td>1.81</td>
<td>71</td>
</tr>
<tr>
<td>Density Peak</td>
<td>Air Quality</td>
<td>22</td>
<td>1.52</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Appliances energy prediction</td>
<td>31</td>
<td>1.71</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>EEG Eye State</td>
<td>32</td>
<td>1.75</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Real-time Election</td>
<td>35</td>
<td>1.74</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Beijing Multi-Site Air-Quality</td>
<td>39</td>
<td>1.77</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Beijing PM2.5</td>
<td>45</td>
<td>1.92</td>
<td>54</td>
</tr>
</tbody>
</table>

4.4. Results on Standard Time Series Data

In this experiment, the evaluation of the space time series clustering is carried out on standard time series databases https://archive.ics.uci.edu/ml/index.php. Table 6 lists the runtime, the ESS value, and the memory usage for different used time series databases. From this table, we can observe that the k-means and DBSCAN are the most powerful methods compared to the other space time series clustering algorithms. Fuzzy c-means and Density Peak are less competitive than k-means and DBSCAN. However, they have obtained reasonable results compared to CHA. This latter is the less competitive algorithm, which requires high computational and memory resources, and it provides less quality of clusters.
4.5. Results on Urban Odense Traffic Data

Figure 3: Quality of Returned Clusters on Urban Odense Traffic Data: ESS

Figure 4: Quality of Returned Clusters on Urban Odense Traffic Data: Classification Ratio
The first experiments aim at comparing the space time series clustering using the large urban Odense traffic data. We have collected several time series in different locations. Each time series constitutes the set of the number of flow values in one hour, where each flow is the number of cars or bikes during a time period. We set the time period to 1min, 2min, 3min, 4min, and 5min, respectively. This allows to create different time series with different sizes (60, 30, 20, 15, 12). We have captured more than 1 million of time series for the experiments and constructed 20 datasets. Each dataset named \( nXmY \) indicates that it contains \( n \) time series with \( m \) different flow values. In the experiments, \( n \) belongs to the set \( \{1,000, 10,000, 100,000, 1,000,000\} \), and \( m \) belongs to the set \( \{12, 15, 20, 30, 60\} \). Figures 3 and 4 showed the quality of the space time series clustering using ESS and classification ratio. Figure 3 showed the ESS value of the different space time series clustering algorithms on dataset 1,000,000X60Y, along with different similarity measures (Euclidean, Manhattan, DTW, STS, Dissim, and PC). We have also observed that by varying the number of clusters in \( k \)-means and fuzzy c-means, respectively, from 1 to 10, epsilon and maximum distance in DBSCAN and hierarchical clustering, respectively from 0.1 to 1.0, DTW provides better results whatever the case used. This is explained by the fact that the DTW measure is well adopted for space time series data by considering both temporal and spatial dimensions of the space time series data. This is why the DTW measure is used for the remaining experiments. Figure 4 presents the classification ratio of different clustering algorithms, and with different Odense locations. The results revealed that the classification ratio of the space time series clustering algorithms is decreased while increasing the number of traffic flow values. Thus, for low density locations, the classification ratio of all algorithms exceeds 80% (\( k \)-means and fuzzy c-means reach 90%). However, for high density locations, the classification ratio goes under 70% for some algorithms such as DBSCAN and CHA. Figure 5, and 6 present the runtime in seconds and the memory usage in mega bytes for the space time series clustering using the 20 datasets of the urban Odense traffic data. By varying the number of space time series data from 1,000 to 1,000,000, and the the number of flow values from 12 to 60, we remark that CHA and c-means are slow. Actually, they require high computational resources to handle the large urban Odense traffic data. This confirms the discussion given in Section 3.4. In general, the space time series clustering algorithms need reasonable memory usage (less than 250 mb), but they require high computational resources (more than 3 hours) for handling 1,000,000 flow
values. We can state that the existing space time series clustering algorithms could handle the large data as the case of urban Odense traffic data.

Figure 5: Runtime of Clustering on Urban Odense Traffic Data

4.6. Results on Urban Beijing Traffic Data

The next experiments aim to show the ability of the space time series clustering algorithms for handling big dataset, as the case of the urban traffic data captured in the second largest city in the world. Figure 7 presents the performance of the space time series clustering algorithms on Beijing locations. When varying the number of flow values from 1 million to 30 million, we have observed that the $k$-means and c-means provide good results in terms of quality of returned clusters for both ESS and classification ratio values. They also are very competitive in terms of runtime and memory usage compared to the other space time series clustering. However, all space time series clustering algorithms are high time consuming for dealing with big dataset; they need several days to group Beijing data having 30 million of flow values. More advanced clustering techniques Schubert et al. 2017;
Figure 6: Memory of Clustering on Urban Odense Traffic Data

Pourkamali-Anaraki and Becker 2017; Q. Zhang et al. 2018; Xiaojun Chen et al. 2018 are needed to be adopted for handling big space time series datasets.

5. Challenges and Future Directions

This section presents some challenges and future applications in space time series clustering.

5.1. Challenges

In this section, we present four challenges in the future work on space time series clustering.

Challenge I: Improving runtime performance of space time series clustering. Space Time series clustering approaches are very time consuming in particular while dealing with many spatial points and huge time series. To handle the big space time series, technologies from different domains could be adapted such as: i) High performance computing (HPC) aims at using parallel frameworks to speed up the sequential solutions Shi et
Figure 7: Results on Beijing Locations
Some of the most well-known architectures apply the multi-core CPU or GPU and perform on MapReduce or Spark platforms; ii) Computational intelligence (CI) is a collection of intelligent methods aiming at optimizing complex problems with strategies like meta-heuristics Queiroga, Subramanian, and Lucido dos Anjos 2018; Y. Djenouri, Drias, and Bendjoudi 2014; Y. Djenouri, Drias, and Habbas 2014; Y. Djenouri and Comuzzi 2017; and iii) Database systems provide techniques to efficiently store, update, and search space time series data, such as query optimization and index optimization. Adapting, combining, and optimizing technologies in space time series also provide many open research questions and future directions.

Challenge II: Improving quality performance of space time series clustering. The quality of the existing space time series clustering approaches became poor for the complex and big time series data. To solve this limitation, the deep network Ni et al. 2018 model can be utilized and applied for handling this situation. In this context, a suitable distribution of the data in the deep network should be considered and performed.

Challenge III: Correlation between space time series data. The existing algorithms for space time series clustering consider individual space time series data and ignore the correlation between the time series data. Studying the correlation between space time series using pattern mining algorithms Yagoubi et al. 2017; Campisano et al. 2018; Y. Djenouri, J. C.-W. Lin, et al. 2019; Y. Djenouri and Comuzzi 2017; Y. Djenouri, Djenouri, et al. 2019 could be helpful for space time series. Such complex or extended systems could be interesting for some excited applications such as urban traffic data Gonzalez et al. 2007.

Challenge IV: Adaptation of advanced and specialized clustering methods Several variants of clustering models could also be adapted to handle the space time series data. Many adaptations to specific scenarios such as spatial data Ng and J. Han 2002; Birant and Kut 2007; J. C.-W. Lin, Y. Li, Fournier-Viger, Y. Djenouri, and J. Zhang 2019, high dimensional data McCallum, Nigam, and Ungar 2000; Rathore et al. 2018; Y. Djenouri, Bendjoudi, et al. 2015; J. C.-W. Lin, Y. Li, Fournier-Viger, Y. Djenouri, and L. S.-L. Wang 2019, time series data Keogh and J. Lin 2005, or streaming data Euán, Ombao, and Ortega 2018; McDowell et al. 2018 remain potential possibilities. All these special scenarios are somehow related to possible scenarios in tackling space time series data and thus studied methods in literature review for these scenarios could also be relevant for adaptations to
5.2. Future Directions

Myriad volumes of space time series data are collected in several applications and domains such as social media, urban traffic, and climate change. In this section, we briefly describe future directions and motivation for applying space time series clustering in different applications and domains.
5.2.1. Social Media Analysis

Social media analysis has received a great attention in World Wide Web. Finding structural properties in a social network such as Twitter is a challenge issue Kwon et al. 2013 in recent decades. Consider the example shown in Figure 8(a) that the tweets mentioned the crude oil prices profile\(^5\). At each location, we can observe that different number of users have accessed this page across several days for evaluation. Applying space time series clustering on these data can discover relevant information, for instance, the number of people located in red countries are approximately the same when the oil petrol changes a lot (highly increased or highly decreased). These could be explained by the fact that the economy of these countries are highly dependent of the oil prices, where people care for a high changes in the price of such natural resources. In general, applying space time series clustering in social media data allows creating user profiling in both spatial and temporal dimensions.

5.2.2. Urban Traffic Flows

Urban traffic data consists of observations like number and speed of cars or other vehicles at certain locations as measured by the deployed sensors. These numbers can be interpreted as traffic flow which in turn relates to the capacity of streets and the demand of the traffic system Belhadi, Y. Djenouri, and J. C.-W. Lin 2019; Y. Djenouri and Zimek 2018; Y. Djenouri, Zimek, and Chiarandini 2018. City planners are interested in studying the impact of various conditions on the traffic flow, leading to finding the correlation between the traffic flows. Consider the example shown in Figure 8(b), it illustrates urban traffic flow of different locations in a given city. For each location, we have observed different flow values represented by a time series. Applying space time series clustering on these data allows to group locations having similar traffic behaviors. For instance, traffic in red locations are quite similar. If the traffic flow increases in \(A\), it increases in \(B\) as well.

5.2.3. Climate Change

Climate change directly effects on the precipitation all over the world. Several research Singh, Lo, and Qin 2017; Karpate et al. 2013 focus on the changes in intensity and frequency of precipitation represented by a time

\(^5\)https://twitter.com/CrudeOilPrices
series. Studying propagation of climate changes around closer locations is a challenge issue. One idea to solve this issue is to exploit space time series clustering. Consider the example shown in Figure 8(c), it presents the temperatures at different locations. At each location, we have observed different temperature values represented by a time series. Applying the space time series clustering on these data can allow us to group locations having similar climate change behaviors. For instance, temperature in red locations are quite similar and if the climate changes in A, it is also changed in B. In this case, we can say that there is a propagation in climate change between A and B.

6. Conclusion

This paper presented an overview on space time series clustering approaches. First, we have discussed three categories of existing clustering approaches, including hierarchical, partitioning, and overlapping space time series clustering. We have also elaborated on how limitations in one case could be beneficial in another case depending on the scenario and available knowledge. Second, we have explained four challenges of space time series clustering and discussed how they might affect the clustering process. Third, we have also provided a case study of existing space time series clustering algorithms on intelligent transportation, targeting to two smart cities (Odense in Denmark with large dataset and Beijing in China with big dataset). We have finally presented a summary of the most relevant directions that could be concluded from the applications of space time series clustering. Overall, whereas solutions to time series clustering has gained high maturity in domains such as image/speech processing, transportation, and bio-medical data; the use of space time series clustering in these domains has become an emerging issue. Our main conclusion from this study is that much exploration and deep progress are still required in all directions to obtain more mature solutions for end-user satisfaction.

References


Gupta, Manish et al. (2014). “Outlier detection for temporal data: A survey”. In: *IEEE Transactions on Knowledge and Data Engineering* 26.9, pp. 2250–2267.


Karypis, George, Eui-Hong Han, and Vipin Kumar (1999). “Chameleon: Hierarchical clustering using dynamic modeling”. In: Computer 32.8, pp. 68–75.


Keogh, Eamonn and Jessica Lin (2005). “Clustering of time-series subsequences is meaningless: implications for previous and future research”. In: Knowledge and information systems 8.2, pp. 154–177.

Koperski, Krzysztof, Junas Adhikary, and Jiawei Han (1996). “Spatial data mining: progress and challenges survey paper”. In: ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 1–10.


Ng, Raymond T. and Jiawei Han (2002). “CLARANS: A method for clustering objects for spatial data mining”. In: IEEE transactions on knowledge and data engineering 14.5, pp. 1003–1016.


Rathore, Punit et al. (2018). “A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data”. In: IEEE Transactions on Knowledge and Data Engineering.


Shi, Xuanhua et al. (2018). “Graph processing on GPUs: A survey”. In: ACM Computing Surveys (CSUR) 50.6, p. 81.


Singh, Saurabh K, Edmond Yat-Man Lo, and Xiaosheng Qin (2017). “Cluster Analysis of Monthly Precipitation over the Western Maritime Continent under Climate Change”. In: Climate 5.4, p. 84.


