BMC Medical Research Methodology

RESEARCH

Open Access

Clustering methods for categorical time series and sequences : a scoping review

Ottavio Khalifa[1*], Alan Balendran[1], Viet-Thi Tran[1,2] and François Petit[1]

Abstract

Objective To provide an overview of clustering methods for categorical time series (CTS), a data structure common in epidemiology, sociology, biology, and marketing, and to support method selection according to data characteristics.

Materials and methods We searched PubMed (via MEDLINE), Web of Science, and Google Scholar up to November 2024 for articles proposing and evaluating CTS clustering techniques. Methods were classified into three families—distance-based, feature-based, and model-based—and assessed for their ability to address challenges such as variable sequence length, multivariate data, continuous time, missing data, covariates, and large data volumes.

Results Of 14,607 records retrieved, 124 articles describing 129 methods were included. Distance-based approaches, especially those using Optimal Matching, were most common, with 56 methods. We found 28 model-based methods, which covered a broader range of complex data structures such as multivariate data, continuous time and time-invariant covariates. We recorded 45 feature-based approaches, which were on average more scalable but less flexible. Fewer than half of the methods provided public implementations. A searchable Web application (https://cts-clustering-scoping-review-7sxqj3sameqvmwkvnzfynz.streamlit.app/) was developed to support method selection.

Discussion CTS clustering methods are highly heterogeneous in assumptions, capabilities, and scalability. Distance-based approaches dominate, but model-based methods offer richer modeling potential, while feature-based ones emphasize performance at the cost of flexibility.

Conclusion This review highlights methodological diversity and gaps in CTS clustering. The proposed typology and Web application aim to help researchers choose appropriate methods to choose appropriate methods for their data.

Keywords Clustering, Categorical, Time Series, Sequence Analysis, Care Trajectories, Review

*Correspondence:
Ottavio Khalifa
ottavio.khalifa@inserm.fr
[1]Université Paris Cité, Université Sorbonne Paris Nord, INSERM, INRAE, Centre for Research in Epidemiology and StatisticS (CRESS), Paris, France
[2]Centre D'Epidemiologie Clinique, AP-HP, Hôpital Hôtel Dieu, Paris F-75004, France

Background

Time series and sequence analysis consider observations of one or multiple variables ordered along a temporal or non-temporal scale. Variables can be real-valued, discrete, ordinal, or categorical. Among these, categorical time series or sequences (CTS) represent a particularly challenging case, as they lack a metric structure and involve symbolic, non-comparable values.

CTS consist of ordered lists of categorical observations and arise in many fields such as epidemiology (care trajectories) [1], sociology (social sequences reflecting socio-economic changes) [2], genomics (DNA sequences) [3], or marketing (clickstream data) [4]. Figure 2 (see Results section) illustrates this general structure.

A central objective of CTS analysis is to identify and group recurrent patterns across individual trajectories, enabling visualization, outlier detection, and hypothesis generation. The proliferation of approaches across disciplines has led to fragmented terminology. For example, in sociology, clustering approaches called "Sequence analysis" were popularized by Andrew Abbott, notably through the use of Optimal Matching (OM) dissimilarities [2, 5], which operationalize sequence comparison via edit distances adapted from alignment algorithms such as the Needleman–Wunsch algorithm, widely used in biology.

Existing reviews on CTS clustering are restricted to specific application domains such as care trajectories [6, 7] or social sequences [8] or to specific method families, and none of them provides a cross-disciplinary synthesis. The present scoping review aims to provide a comprehensive, cross-disciplinary synthesis of CTS clustering methods with the objective of guiding method selection according to key data characteristics, including sequence length heterogeneity, multidimensional observations, irregular or continuous sampling, missing data mechanisms, incorporation of time-invariant covariates, and scalability to large CTS datasets

Methods

This review follows was reported according to the PRISMA-ScR guidelines [9].

Search strategy

We searched Google Scholar, Web of Science, and PubMed (MEDLINE) from inception to November 11, 2024 using keywords related to "clustering", "categorical", and "sequence". To capture domain-specific approaches, expressions such as "Sequence Analysis" were included. Full search terms are provided in the Appendix. Our preliminary search revealed that each community refers to the CTS clustering task using different terminology; we therefore combined keywords related to the objects being clustered (e.g., "care trajectories", "DNA sequences") with keywords related to clustering methods (e.g., "latent class", "dissimilarity"), and used synonyms for "categorical" such as "nominal".

Eligibility criteria

We included articles proposing CTS clustering methods and testing them on synthetic or real datasets. Studies restricted to numerical or ordinal sequences were excluded. Purely univariate binary sequences were also excluded because they form a distinct literature (e.g., time series of 0/1 events) that does not generalize to the CTS setting we focus on. However, methods applicable to binary multivariate sequences — which can represent one-hot encoded categorical sequences — were included. We also excluded: methods without clustering evaluation, clustering-by-example approaches, clustering timestamps instead of sequences [10], pure application papers with no methodological description, and non-English or abstract-free publications.

Study selection

Screening was conducted in Covidence. Two reviewers (OK and FP) first screened 100 records to align criteria, then the remaining titles/abstracts were reviewed by one reviewer (OK). Next, 142 studies were screened by two reviewers (OK and FP), and the remainder by one reviewer (OK).

Data extraction

One reviewer (OK) extracted CTS clustering methods from included articles. When an included article described several methods, for instance, proposing a new method while also using older ones as baseline comparators, the original paper introducing each method was retrieved and used as the primary reference. Supplementary materials and online code were also checked.

A second reviewer (AB) independently checked 10 studies for data extraction.

We collected characteristics of datasets used to evaluate methods: number of sequences (N), number of categorical variables (m), numbers of levels $s_1, \ldots, s_m$, and longest sequence length (T). For multiple datasets, we retained the one with the largest value of $N \times |S_k| \times T$.

We noted the availability of code/software, how the number of clusters was determined (automatic, fixed by the user, or using post-hoc metrics such as Silhouette [11] or BIC [12]), and the number of hyperparameters.

Analysis

Communities and use cases

Each method was assigned to a community based on the disciplinary category of the publishing journal (Web of Science) or on the first author's academic affiliation (stated in the title of the article) for preprints. We

also described the data used to apply each method and grouped them into eight "use cases" (see Fig. 2 in Description section).

Clustering approaches

CTS clustering involves two subtasks: representing categorical sequential data, and applying a clustering algorithm. The first subtask consists in mapping CTS into a suitable representation — a distance matrix, a feature vector, or a latent model — on which the clustering algorithm is then applied in the second step. Three main approaches are commonly used for this first step, defining three method families:

- Distance-based: define dissimilarity measures for CTS and apply clustering algorithms to the resulting distance matrix (e.g., hierarchical clustering, k-medoids). Optimal Matching (OM), also called sequence alignment, edit or Levenshtein distance, is a key example. Here, "distance" and "dissimilarity" are used interchangeably, although these functions may violate metric properties such as triangle inequality or symmetry. Any method defining a dissimilarity function on CTS and processing a distance matrix is considered distance-based.
- Model-based: define generative models for CTS (e.g., Markov Chains, Generalized Linear Models) and estimate them on the dataset. Clustering is then derived from the learned model. Any method that fits a generative model to the entire dataset and uses it for clustering is considered model-based.
- Feature-based methods transform categorical time series into numerical vectors, allowing to process with standard clustering methods. This transformation results in loss of information. Features can be constructed by counting subsequences [13], estimating empirical transition matrices on each sequence [14], or computing autocorrelation vectors [15]. Once sequences are transformed into vectors of dimension d, they can be processed using generic clustering algorithms such as k-means or DBSCAN with any dissimilarity function on $\mathbb{R}^d$. Any method that embeds sequences into a vector space is considered feature-based.

Each method was classified according to its family, then grouped into subcategories based on the distance, model, or featurization used. A complete list with discussion is provided in the Appendix.

Once the representation (distance, model, or features) is set, a clustering paradigm is applied. We identified the main clustering paradigm used in this second step, such as partition-based (e.g., k-means), hierarchical (agglomerative/divisive), density-based (e.g., DBSCAN), and model-based approaches.

Accounting for specific data structures

We examined whether each method: 1) handled varying sequence lengths; 2) supported multivariate CTS (i.e. multiple categorical variables observed over time); 3) treated time as a continuous variable rather than discretized; 4) included a mechanism for handling missing values (e.g. imputation); 5) incorporated time-invariant covariates (e.g., age, gender) in the clustering process; and 6) scaled to large datasets (defined as $N \times |S_k| \times T > 10^7$). The first four characteristics were defined prior to the search from expert knowledge; time-invariant covariates and scalability emerged from our reading of the literature.

Each characteristic reflects a common challenge in CTS analysis: multivariate data and variable-length sequences arise naturally in many applications [16]; irregularly sampled observations, common in observational health data, naturally call for treating time as a continuous variable rather than a discrete index [17]; missing data are common in longitudinal studies, where participants may drop out or miss visits [18]; time-invariant covariates such as age or sex are routinely collected alongside sequence data, especially for health applications [19]; and scalability becomes critical when datasets are large.

Dependency order

We assessed the dependency order each method can capture, defined as the maximum transition order considered when discriminating between sequences. If $d = 0$, the method does not take temporal information into account and is agnostic to permutations of columns on the dataset. This is the case for Hamming distance and basic GBTM approaches. If $d = k > 0$, the method considers transitions of orders 1 to k, and its clustering function can be entirely built from $k+1$-grams — this is the case for Markov-based approaches, with $k = 1$ in most cases. If $d = \infty$, the method cannot be built from k-grams for any finite k and considers transitions of any order — this is the case for OM-based approaches, which can align subsequences of arbitrary length. Further details are provided in the Appendix.

Results

Description

The electronic search retrieved 14607 unique articles. After full-text review, 124 articles were included, from which we identified 129 methods. These methods were used in eight main communities. The most represented was Artificial Intelligence with 30 (23.3%) methods, followed by Biology, Social Science (sociology and

economics), Statistics, Computer Science, and Healthcare. Figure 1 summarizes the screening process.

Full agreement was found at both stages involving two independent reviewers — title/abstract screening and data extraction.

Methods were evaluated in 171 applications, including biological sequences (DNA, protein), social sequences, care trajectories, and clickstream data. Figure 2 summarizes the main use cases for CTS clustering, with the number of occurrences found in the literature and illustrated examples. Table 6, mapping methods to communities, is provided in the Appendix.

Half ($n = 70, 54.3\%$) of the 129 identified methods did not provide a public implementation of any form. Among the 58 other methods, we differentiated the proposed methods with a simple publicly available code (typically via a GitHub link) ($n = 30, 23.3\%$) from those implemented in a package ($n = 22, 17.1\%$) or free software ($n = 6, 4.7\%$).

Identified methods and their characteristics

Tables 1, 2 and 3 summarize all the information we extracted for distance-based, feature-based and model-based methods respectively. Due to space constraints, only the methods provided with public code are presented. Complete versions of these tables are provided in Appendix (section 8.7), adding methods from references 94-152 an overview of methods by family, domain, and supported data characteristics can also be explored interactively in our Web Application[1], which allows filtering by any combination of these criteria.

Distance-based methods

We identified $n = 56$ (43.4%) distance-based methods, including $n = 40$ (71.4%) derived from Optimal Matching (OM) dissimilarity or its simplified forms (e.g., Hamming, Jaro, Longest Common Subsequence). Several studies, including those cited in [8], propose new strategies to tune OM hyperparameters, particularly substitution costs [36, 43].

Among these methods, $n = 24$ (42.9%) rely on hierarchical clustering,with a marked preference for Ward linkage ($n = 8$). Partitioning algorithms such as *k*-medoids appear in $n = 9$ methods (16.1%), and graph-based approaches in $n = 6$ (10.7%), typically through partitioning on neighborhood graphs. Most ($n = 50$, 89.3%) handle varying sequence lengths. In particular, OM dissimilarity allows deletion, enabling comparison of sequences of different lengths, although preprocessing to equalize lengths is often recommended (see the chapter of [78] focusing on Care Trajectories analysis).

Multivariate CTS are supported by $n = 9$ (16.1%) distance-based methods. The most used strategy is Multichannel Sequence Analysis (MSA), extending OM to multiple variables [41]. Another option is encoding multivariate data in an Extended Alphabet. See [40] for a comparison of these two approaches. Continuous time is considered by $n = 5$ (8.9%) methods, through adaptations of OM, Hamming, or Dynamic Time Warping [17, 20, 79].

Missing values are explicitly handled in $n = 5$ (8.9%) methods. The only documented approach is to consider "Missing" as an additional categorical state, despite risks of artificial cluster formation as documented in [18]. Only one recorded distance-based approach integrates time-invariant covariates [19], with an application on care trajectories.

The computational cost of the OM dissimilarity is high and has been widely addressed in bioinformatics. As a result, classical OM-based methods are generally applied to relatively small datasets, with the product $N \times S \times T$ ranging from 10^5 to 10^7. To improve scalability, several greedy heuristics avoid full pairwise distance computation [26, 33, 39], enabling $N \times S \times T > 10^{10}$ and up to 10^{12}. These are referred to as selective distance-based methods.

Among distance-based methods, $n = 11$ (19.6%) have no reported hyperparameters. The most common are weights, especially substitution cost schemes (one weight per pair of categorical attributes) of OM [8], reported in $n = 10$ (18.0%) methods. These may be expert-driven (e.g., for DNA sequences) or data-driven (see [8] for an overview). Weights are also used in mixed Euclidean distances [80]. Finally, $n = 6$ (10.7%) methods require a similarity threshold (e.g., greedy [39] or graph-based approaches [81]).

Feature-based methods

We recorded $n = 45$ (34.9%) feature-based methods, showing a wide variery of strategies to construct feature vectors from CTS. The most common is using *k*-mers (or *n*-grams), found in $n = 13$ (28.9%) methods, where sequences are represented by subsequences of fixed length. These vectors are clustered directly [13, 82] or used in self-supervised learning pipelines [52, 54]. Other featurization techniques include Chaos Game Representation [58], empirical transition matrices [14], Fourier transforms [51], or latent representations obtained from deep learning [60].

Once vectorized, CTS can be clustered with standard numerical algorithms. Among the 45 approaches, partition-based methods (e.g., *k*-means, *k*-medoids, bisecting *k*-means) are most frequent ($n = 21, 46.7\%$), followed by hierarchical clustering ($n = 17, 37.8\%$).

[1] https://cts-clustering-scoping-review-7sxqj3sameqvmwkvnzfynz.streamlit.app/

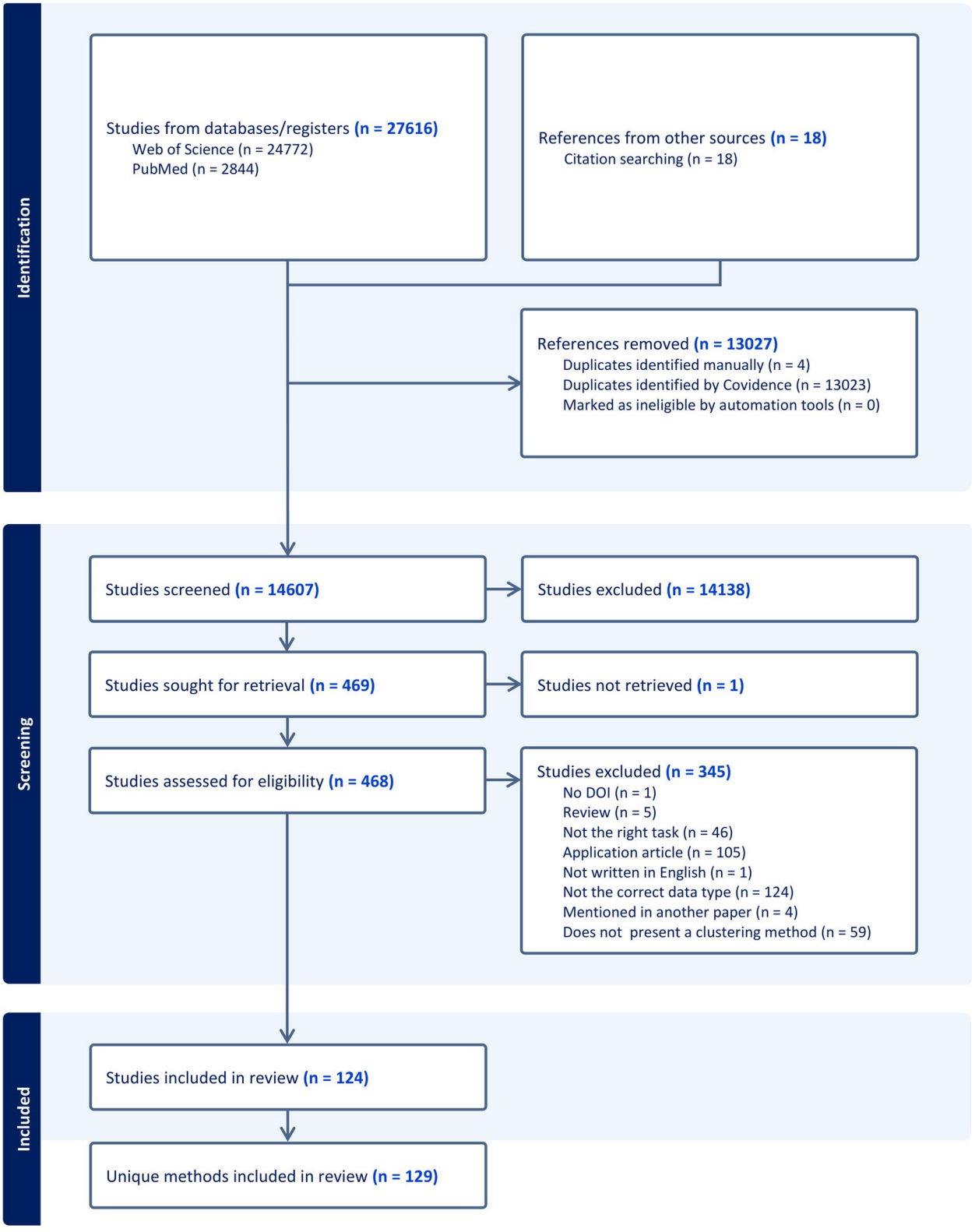


Fig. 1 Prisma flow diagram

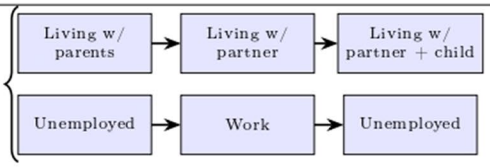
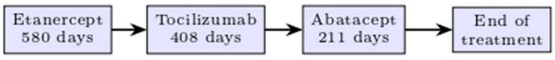
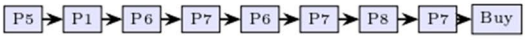
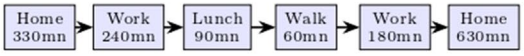
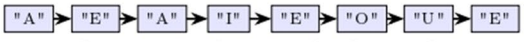
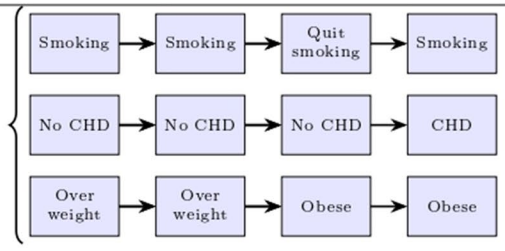
Use Case	Number	Example from the literature	Ref
Biological Sequences (DNA, Protein)	52 (30.4%)	 DNA Sequence.	[50]
Social sequences	31 (18.1%)	 Bivariate Work and Family Trajectory.	[37]
Care trajectories	19 (11.1%)	 Continuous Time Care trajectory for rhumatoid artritis patient (source : [97]).	[38]
Clickstream Data	14 (8.2%)	 Sequence of user clicks on products (P) from an online store.	[15]
Activities Sequence	11 (6.4%)	 Continuous time sequence of daily activities.	[17]
Speech sequences	6 (3.5%)	 Sequence of recognized vowels in a speech.	[98]
Clinical trajectories	4 (2.3%)	 Sequence of multiple clinical categorical outcomes of a patient.	[99]

Fig. 2 Main use cases for CTS clustering, with number of found occurrences and illustrated examples from the literature

All 45 methods support varying sequence lengths. However, most originate from bioinformatics and focus on DNA, RNA, or protein sequences, leaving some structures underexplored. Continuous time is handled in $n = 2$ (4.4%) methods [53, 56], and $n = 4$ (8.9%) process multivariate CTS [59, 60, 83, 84]. Only one addresses missing data [14], and none incorporate time-invariant covariates.

Many of these methods have been developed in bioinformatics to process large datasets without computing

Table 1 Distance-based methods with public code and their characteristics

Name	Subfamily	Dependency order	Continuous time	Various lengths	Missing data	Multi variable	Covariates	Main data type	Data volume			Ref.
									N	T	S	
Hamming Distance	Hamming	0	×	×	×	×	×	Activities	1200	1400	7	[20, 21]
Dynamic Hamming Distance	Hamming	1	×	×	×	×	×	Social	8000	150	2	[8, 22]
Fuzzy Temporal Hamming Distance	Hamming	0	✓	×	×	×	×	Activities	1200	1400	7	[20]
Damerau – Levenshtein Distance	OM	User	×	✓	×	×	×	Care Trajectories	1000	15	15	[23, 24]
Modified Needleman-Wunsch	OM	∞	×	✓	×	×	×	Care trajectories	1000	15	15	[23]
Jaro Distance	OM	User	×	✓	×	×	×	Care trajectories	1000	15	15	[23, 25]
Clover	Selective Hamming	User	×	✓	×	×	×	DNA	10^{10}	150	4	[26]
Randomized Discrete Sequence Clustering	Random search	∞	×	✓	×	×	×	Clickstream	5000	7000	3×10^4	[27]
Tribe-MCL	Selective OM	∞	×	✓	×	×	×	Protein	10^5	400	20	[28]
Contextual Edit Distance	OM	∞	×	✓	×	×	×	Activities	1200	1400	7	[29]
GClust	LCS	∞	×	✓	×	×	×	DNA	10^5	10^7	4	[30]
Laplacian Eigenmaps	OM	∞	×	✓	×	×	×	DNA	30	500	4	[31]
ProClust	OM (local)	∞	×	✓	×	×	×	Protein	6×10^4	350	20	[32]
Cd-hit	Selective OM	∞	×	✓	×	×	×	Protein	6×10^4	500	4	[33]
Slymfast	Random search	User	×	✓	×	×	×	DNA	2×10^7	100	4	[34]
Jaro-Winkler Distance	OM	User	×	✓	×	×	×	Care trajectories	1000	15	15	[23, 35]
Optimal Matching	OM	∞	×	✓	×	×	×	Social	300	50	35	[5, 8]
OMsloc	OM	∞	×	✓	×	×	×	Social	700	100	4	[8, 36]
Longest Common Subsequence	LCS	∞	×	✓	×	×	×	DNA	7×10^6	75	4	[37]
LinCLust	Selective OM	∞	×	✓	×	×	×	Protein	10^8	400	20	[38]
BlastClust	Selective OM	∞	×	✓	×	×	×	DNA	10^5	10^7	4	[30, 39]
Extended Alphabet	OM	∞	×	✓	×	✓	×	Social	1700	26	$3 \times 5 \times 4 \times 2$	[40]
Multichannel Sequence Analysis	OM	∞	×	✓	×	✓	×	Social	1800	50	$7 \times 12 \times 10$	[40, 41]
AliClu	OM	∞	✓	✓	×	×	×	Care trajectories	426	20	8	[17]
GIMSA	OM	∞	×	✓	×	✓	×	Social	1400	50	4×5	[42]
OMstran	OM	∞	×	✓	✓	×	×	Social	2400	20	4	[8, 43]

Table 2 Feature-based methods with public code and their characteristics

Name	Subfamily	Dependency order	Continuous time	Various lengths	Missing data	Multivariate	Covariates	Data type	Data volume			Ref.
									N	T	S	
nTreeClus	Self-supervised (Random Forest)	User	×	✓	×	×	×	Social	20000	90	20	[44]
mBKM	Entropy	User	×	✓	×	×	×	DNA	4×10^4	100	4	[45]
ISCT (Interpretable Sequence Clustering Tree)	kmers	∞	×	✓	×	×	×	Clickstream	5000	7000	3×10^4	[46]
iDeLUCS	Chaos Game Representation	User	×	✓	×	×	×	DNA	3200	5×10^4		[47]
Sqn2Vec	Self-supervised learning (latent space)	User	×	✓	×	×	×	Other	5000	3×10^4	7000	[48]
DeLUCS	Chaos Game Representation	User	×	✓	×	×	×	DNA	1200	5×10^4		[49]
STARS	Fourier	∞	×	✓	×	×	×	DNA	30	3.4×10^6		[50]
Envclust	Fourier	∞	×	✓	×	×	×	Physiological	80	1000	6	[51]
MeShClust (v 3.0)	kmers	User	×	✓	×	×	×	DNA	10^4	4×10^6		[52]
NMS (Number of matching subsequences)	SVR	∞	×	✓	×	×	×	Social	4300	15	8	[53]
MeShClust (v 1.0)	kmers	User	×	✓	×	×	×	DNA	100	10^4	4	[54]
Categorical Functional Data Analysis (CFDA)	Functional data analysis	∞	✓	×	×	×	×	Care	3000	4	10	[55, 56]
ctsfeatures	Correlation	User	×	✓	×	×	×	Protein	20	7000	4	[15, 57]
kmer	kmers	User	×	✓	×	×	×	DNA	**NEI**	**NEI**	**NEI**	[13]
CGR	Chaos Game Representation	∞	×	✓	×	×	×	Protein	**NEI**	**NEI**	**NEI**	[58]
cluster Clickstreams	Markov	User	×	✓	✓	×	×	Clickstream	10^5	50	7	[14]
Transition Matrix features	Markov	1	×	✓	×	✓	×	Social	120	90	(2×2)	[59]
SVR (Subsequence Vector Representation)	SVR	∞	✓	✓	×	×	×	Social	4300	15	8	[53]
User Journey2Vector	Self-supervised learning (latent space)	User	×	✓	×	✓	×	Clickstream	**NEI**	**NEI**	**NEI**	[60]

Table 3 Model-based methods with public code and their characteristics

Name	Subfamily	Dependency order	Continuous time	Various lengths	Missing data	Multivariable	Covariates	Main data type	Data volume			Ref.
									N	T	S	
Markov Chain Clustering	Markov	1	×	✓	×	×	×	Social	10000	30	6	[61, 62]
Dirichlet Multinomial Clustering	Markov	∞	×	✓	×	×	×	Social	10000	30	6	[61]
Clickclust	Markov	1	×	✓	×	×	×	Clickstream	320	360	17	[63, 64]
DBHC	HMM	1	×	✓	×	×	×	Social	2000	20	8	[65]
MedSeq	Exponential	0	×	×	×	×	✓	Social	700	70	6	[66]
LCA (Latent Class Analysis)	GLM	0	×	×	✓	✓	×	Social	2290	6	32	[67, 68]
Look-A-Liker (LAL)	RNN	∞	✓	✓	×	×	×	Clickstream	10^6	50	5	[69]
GBTM (Group-based Multi-Trajectory Modelling)	GLM	0	×	×	×	✓	✓	Care trajectories	10^6	16	2×8	[70, 71]
TESS (Temporal Event Sequence Summarization)	Combinatorial	∞	✓	✓	×	✓	×	Other	200	50000	12	[72]
DMHP (Dirichlet Mixture of Hawkes Processes)	Hawkes	∞	✓	✓	✓	×	×	Care trajectories	2000	50	5	[73]
Mixture Hidden Markov Models	HMM	1	×	✓	✓	✓	✓	Social	2000	16	3×2	[74]
lamb (Latent factor Mixture model for Bayesian clustering)	GLM	∞	✓	✓	×	✓	✓	Physiological	300	1000	10×2	[75]
Linear mixed-effect model	GLM	∞	✓	✓	✓	×	✓	Social	20000	4	2	[76]
Class-specific GLM	GLM	∞	✓	✓	✓	×	✓	Physiological	750	126	3	[77]

all pairwise dissimilarities. Aside from the 'selective' distance-based methods mentioned earlier, feature-based methods are those that allow for handling the largest data volumes, particularly in the field of bioinformatics. Among them, $n = 8$ (17.8%) were tested on large-scale datasets, with $N \times S \times T$ ranging from 10^8 to 10^{11}.

Their most common hyperparameter encodes the dependency order (Appendix), reported in $n = 20$ (44.4%) methods. It corresponds to subsequence length in *k*-mers [13], sliding window size [44], or Markov order [14], and governs both complexity and runtime. Additionally, $n = 9$ (20%) rely on ML algorithms, requiring hyperparameters such as learning rate, number of layers for deep learning approaches [60], or number of trees for Random Forest approaches [44].

Model-based methods

We identified $n = 28$ model-based methods (21.7%). Markov chains and Hidden Markov Models (HMMs) and variants dominate ($n = 16$, 57.1%). Generalized Linear Models (GLMs) adapted to CTS data appear in $n = 6$ approaches (21.4%). The Expectation–Maximization (EM) algorithm is used in $n = 17$ methods (60.7%) for mixture estimation, while Bayesian approaches rely on Markov Chain Monte Carlo in $n = 7$ methods (25%).

Most model-based methods handle varying sequence lengths ($n = 25$, 89.3%). A subset supports multivariate CTS ($n = 9$, 32.1%), including GLM [71, 75] and HMM [74, 85] extensions, and a Transformer-based method [60].

Continuous time is addressed in $n = 10$ methods (35.7%), through adaptations of Markov models [86, 87], time-dependent GLMs [77, 88], or latent functional representations [56, 69]. Missing data are explicitly considered in $n = 10$ methods (35.7%), generally by model-based imputation. Time-invariant covariates are incorporated in $n = 11$ approaches (39.3%), typically as predictors of cluster membership [62] via a multinomial logit model, or fixed effects in GLMs [75].

Model-based approaches are notoriously less scalable than other families due to the computational complexity of algorithms such as EM or MCMC. As a result, only $n = 4$ methods have been tested on large datasets (14.3%), while most target $N \times S \times T$ from 10^4 to 10^7.

All of the $n = 17$ methods relying on the EM algorithm require initialization. In practice, this is often done via grid search. Bayesian approaches require prior specification. One Markov-based method leaves dependency order open [89], while others fix it to order 1. GLM-based methods require defining model structure (e.g., time effects, random effects) and selecting a baseline category for identifiability.

Number of clusters choice

Among the 129 methods, $n = 55$ (42.6%) did not provide a generic way to select the number of clusters, leaving the choice to the user. $n = 19$ methods (14.7%) determine it automatically (e.g., via stopping rules). The remaining approaches rely on post-hoc criteria comparing multiple solutions. AIC or BIC are used in $n = 18$ methods $(14.0\%$), mainly within model-based approaches, as they balance likelihood and complexity. Among distance-based and feature-based methods, Silhouette-based criteria [11] are most common, including Average Silhouette Width (ASW), used in $n = 13$ methods. Other criteria appear sporadically without a clear dominant trend.

Web application

To make navigation through our results as easy as possible, we built a Streamlit Web Application[2]. This application allows to filter methods based on a combination of specific cases (e.g : covariates and multivariate) or on tested use-cases (e.g if the user only wants methods that have been already tested on care trajectories). Its code and the list of methods are open source and can be improved based on readers suggestions.

Discussion

Our scoping review identified 129 methods for CTS clustering. The variety of approaches stems from two main challenges in CTS analysis: giving a metric structure to categorical data, and modeling dependency. From these two challenges, three types of approaches emerged from the literature: distance-based, feature-based and model-based CTS clustering, each with several subcategories.

A comparison of the three method families for CTS clustering

Distance-based approaches remain the most common, with OM being the most popular overall. Their main advantage is that they are non-parametric and require no assumption on the generative process of the data. A wide variety of distances is available, among which OM and its variants allow researchers to finely define and interpret the notion of dissimilarity between sequences for their specific use case; once this is done, applying a well-established clustering algorithm such as hierarchical Ward or k-medoids is straightforward. Reliable and well-documented software such as TraMineR [90] also facilitates their use. However, distance-based methods have notable limitations: the choice of substitution costs in OM remains difficult and debated [8]; scalability is limited for classical OM-based methods, though selective heuristics partially address this [26, 33, 39]. Missing data are poorly handled among distance-based methods, with treating missingness as an additional state being the only documented strategy, despite its tendency to create artificial clusters as reported in [18].

Among the three families, model-based methods offer the best ability to handle specific data structures. Their probabilistic framework yields soft cluster assignments and enables uncertainty quantification; confidence intervals on model parameters can be derived, and the fitted model can also be used as a data generator, which is useful for simulation or data augmentation. GLM and HMM-based approaches are particularly flexible, and [88] even addresses all five considered challenges simultaneously. Reliable implementations are available, notably seqHMM [74] for hidden Markov model-based clustering. However, these approaches are also the most demanding: they are harder to use, involve more hyperparameters (e.g. prior choice for Bayesian approaches, choice of a baseline categorical state for GLMs, EM initialization), and rely on stronger assumptions about the generative process of the data. They also scale less easily, although some methods have been tested on moderately large datasets [63, 76, 88].

At the other end, feature-based approaches are the most scalable, with several methods tested on datasets of $N \times S \times T > 10^8$. Embedding sequences into a vector space enables the use of standard clustering algorithms, as well as dimensionality reduction and visualization (e.g. PCA, t-SNE) of the cluster structure. They present a large variability in how feature vectors are constructed and processed, with *k*-mers being the most popular approach; *k*-mers also carry a natural Markovian interpretation, and integrate well with modern self-supervised learning pipelines. Many of them were developed to avoid computing all alignment distances; for this reason, they are often referred to as "alignment-free" methods [91], especially in large-scale applications. However, feature-based methods remain the least flexible family: most were developed for biological sequences and leave several structural cases — continuous time, missing data, covariates — unaddressed, particularly when they appear simultaneously. A feature-based approach that could simultaneously address several of these challenges while remaining scalable seems a promising direction for future research.

Choosing a method based on dataset characteristics

To pick a CTS clustering method, researchers should consider six data characteristics: varying sequence lengths, multivariate data, continuous time, missing data, time-invariant covariates, and large data volumes. These issues are especially common in epidemiology with observational datasets presenting irregular sampling, baseline data, missingness, and censoring.

[2] https://cts-clustering-scoping-review-7sxqj3sameqvmwkvnzfynz.streamlit.app/

If sequences have variable lengths, most methods are applicable, as this is the best-covered characteristic ($n = 118$, 91.5%). For multivariate CTS, the choice is more constrained ($n = 23$, 17.8%) : Multichannel Sequence Analysis [41] and Extended Alphabet [40] are the main distance-based options, while GLM [71, 75] and HMM [74] extensions cover the model-based side. When time is continuous or irregularly sampled, only $n = 18$ methods (14.0%) are applicable, mostly model-based or specific adaptations of OM and Hamming [17, 79].

Missing data remains the least covered characteristic ($n = 16$, 12.4%). In practice, model-based approaches are the most principled option, as they can perform model-based imputation [74]; the only documented strategy in distance-based methods - treating missingness as an additional state - carries risks of artificial cluster formation [18]. When time-invariant covariates must be incorporated, model-based approaches are again the natural choice [74, 75, 77, 88, 92], with the exception of one distance-based method [19].

For large datasets, feature-based and selective distance-based methods are the most appropriate, as they avoid computing all pairwise dissimilarities. Most of these were developed in the bioinformatics field, with $N \times S \times T$ ranging from 10^8 to 10^{11}. Finally, the availability of public code should also guide method selection : half of the identified methods ($n = 70$, 54.3%) do not provide a public implementation, which creates a significant burden for researchers who would need to reimplement them.

Limitations

Our review has some limitations. First, even though the number of identified methods (129) can be considered high, we had to make choices in the keyword search that may have caused us to miss a few, or even entire fields of application for CTS clustering. Additionally, most screening and data extraction were performed by a single reviewer, which may introduce a risk of missed studies or inconsistent coding decisions, despite joint screening at key stages and full agreement observed during double extraction.

Second, it does not seem reasonable to benchmark all of the 129 methods at a time, since we showed that they are not designed for the same use-cases. Moreover, even benchmarking a few methods inside a specific use-case requires choosing several generative mechanisms to build synthetic datasets and assessing each clustering method knowing the ground truth. Otherwise, a method may be artificially advantaged if the evaluation relies primarily on mechanisms that closely match its underlying assumptions. Finally, one or several real-life datasets are required, for example to perform a sensitivity analysis. This falls outside the scope of this review, but could constitute another work in its own right.

Thirdly, our choice of $N \times |S| \times T$ as a simple proxy for dataset size is debatable; most reviewed papers do not report runtime, memory usage, or hardware specifications, making a more precise scalability comparison infeasible.

Finally, as we focused exclusively on the clustering task, we overlooked many CTS analysis tools and models, since they did not provide clusters natively.

Open problems and future research directions

Several challenges identified in this review remain underexplored and call for future work.

Missing data handling is the least covered characteristic among the reviewed methods; beyond classical imputation strategies, some forms of missingness such as right-censored sequences — common in survival and epidemiological data — remain largely unexplored in the CTS clustering literature. A dedicated benchmark comparing available strategies across CTS clustering methods would be a valuable contribution, building on recent work by [93].

More broadly, a systematic benchmark comparing clustering methods across families, using synthetic datasets generated from both HMM and GLM frameworks, under multiple scenarios varying the degree of non-stationarity, alphabet size, sequence length, and dataset size, would make a significant contribution to the CTS literature.

Scalability also remains an open problem outside of bioinformatics: most methods applicable to health or social data do not scale beyond $N \times S \times T \sim 10^7$, and computational optimization of existing software such as parallelization of TraMineR [90] or seqHMM [74], would significantly improve their practical usability on larger datasets.

Feature-based methods in particular could benefit from extensions addressing multivariate data, missing data, and covariates simultaneously, while remaining scalable and geometrically interpretable, which none of the reviewed methods currently achieves.

Finally, reproducibility remains a concern : researchers looking for reliable implementations should note that TraMineR [90] provides a comprehensive suite for distance-based methods; seqHMM [74] is the reference implementation for mixture hidden Markov model-based clustering; and ctsfeatures [57] offers feature-based clustering tools for CTS. We encourage authors of future methods to provide at minimum a public code repository alongside their publication.

Conclusion

This scoping review identified 129 methods for CTS clustering across 8 communities, organized into three families and assessed along six data characteristics. Key gaps remain, particularly in handling missing data, scalability

outside bioinformatics, and flexibility of feature-based methods. The companion Web Application aims to help researchers navigate this landscape and select the most appropriate method for their data.

Abbreviations

CTS	Categorical Time Series
OM	Optimal Matching
GLM	Generalized Linear Models
SVR	Subsequence Vector Representation
RNN	Recurrent Neural Network
HMM	Hidden Markov Models
EM	Expectation-Maximization (algorithm)

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-026-02857-6.

Additional file 1. Supplementary Material 1 [94–152].

Acknowledgements
The authors have no acknowledgments to declare.

Authors' contributions
OK, FP and VTT conceptualized this scoping review. OK and FP screened the articles. OK, FP and AB participated in the full-text data extraction. OK and VTT drafted the manuscript. All authors participated in revisions of the manuscript.

Funding
This work was supported by the French Agence Nationale de la Recherche (ANR) through the project reference ANR-22-CPJ1-0047-01.

Data availability
All extracted data was stored in an Excel file, available on this GitHub page: https://github.com/OttaKhalifa/CTS-Clustering-Scoping-Review.

Declarations

Competing interests
The authors declare no competing interests.

Received: 21 November 2025 / Accepted: 20 April 2026
Published online: 28 May 2026

References

1. Vanasse A, Courteau M, Ethier JF. The '6W'multidimensional model of care trajectories for patients with chronic ambulatory care sensitive conditions and hospital readmissions. Public Health. 2018;157:53–61.
2. Liao TF, Bolano D, Brzinsky-Fay C, Cornwell B, Fasang AE, Helske S, et al. Sequence analysis: its past, present, and future. Soc Sci Res. 2022;107:102772. https://doi.org/10.1016/j.ssresearch.2022.102772.
3. Posada D. Bioinformatics for DNA sequence analysis. Springer; 2009.
4. Montgomery AL, Li S, Srinivasan K, Liechty JC. Modeling online browsing and path analysis using clickstream data. Mark Sci. 2004;23(4):579–95.
5. Abbott A, Hrycak A. Measuring resemblance in sequence data: an optimal matching analysis of musicians' careers. Am J Sociol. 1990;96(1):144–85.
6. Mathew S, Peat G, Parry E, Sokhal BS, Yu D. Applying sequence analysis to uncover 'real-world' clinical pathways from routinely collected data: a systematic review. J Clin Epidemiol. 2024;166:111226. https://doi.org/10.1016/j.jclinepi.2023.111226.
7. Flothow A, Novelli A, Sundmacher L. Analytical methods for identifying sequences of utilization in health data: a scoping review. BMC Med Res Methodol. 2023. https://doi.org/10.1186/s12874-023-02019-y.
8. Studer M, Ritschard G. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. J R Stat Soc Ser A Stat Soc. 2015;179(2):481–511. https://doi.org/10.1111/rssa.12125.
9. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc. 2015;13(3):141–6. https://doi.org/10.1097/xeb.0000000000000050.
10. Sangma JW, Sarkar M, Pal V, Agrawal A, Yogita. Hierarchical clustering for multiple nominal data streams with evolving behaviour. Complex Intell Syst. 2022;8(2):1737–61. https://doi.org/10.1007/s40747-021-00634-0.
11. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
12. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978. https://doi.org/10.1214/aos/1176344136.
13. Wilkinson S. kmer: Fast K-Mer counting and clustering for biological sequence analysis. 2023. R package version 1.1.1. https://cran.r-project.org/web/packages/kmer/index.html. Accessed : November 2024.
14. Scholz M. R Package clickstream: Analyzing Clickstream Data with Markov Chains. J Stat Softw. 2016;74(4). https://doi.org/10.18637/jss.v074.i04.
15. Lopez-Oriona A, Vilar JA, D'Urso P. Hard and soft clustering of categorical time series based on two novel distances with an application to biological sequences. Inf Sci. 2023;624:467–92. https://doi.org/10.1016/j.ins.2022.12.065.
16. Istvan M, Duval M, Hodel K, Aquizerate A, Chaslerie A, Artarit P, et al. Evolution of the profiles of new psychotropic drug users before and during the COVID-19 crisis: an original longitudinal approach through multichannel sequence analysis using the French health-care database. Eur Arch Psychiatry Clin Neurosci. 2024;(1):12. https://doi.org/10.1007/s00406-024-01774-3.
17. Rama K, Canhão H, Carvalho AM, Vinga S. AliClu - temporal sequence alignment for clustering longitudinal clinical data. BMC Med Inform Decis Mak. 2019. https://doi.org/10.1186/s12911-019-1013-7.
18. Lazar A, Jin L, Spurlock CA, Wu K, Sim A. Data quality challenges with missing values and mixed types in joint sequence analysis. In: 2017 IEEE International Conference on Big Data (Big Data). IEEE; 2017. pp. 2620–2627.
19. Le Meur N, Vigneau C, Lefort M, Lebbah S, Jais JP, Daugas E, et al. Categorical state sequence analysis and regression tree to identify determinants of care trajectory in chronic disease: Example of end-stage renal disease. Stat Methods Med Res. 2019;28(6):1731–40.
20. Moreau C, Devogele T, de Runz C, Peralta V, Moreau E, Etienne L. A Fuzzy Generalisation of the Hamming Distance for Temporal Sequences. In: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2021. pp. 1–8. https://doi.org/10.1109/FUZZ45933.2021.9494445.
21. Hamming RW. Error detecting and error correcting codes. Bell Syst Tech J. 1950;29(2):147–60. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.
22. Lesnard L. Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. Sociol Methods Res. 2010;38(3):389–419. https://doi.org/10.1177/0049124110362526.
23. Aspland E, Harper PR, Gartner D, Webb P, Barrett-Lee P. Modified Needleman-Wunsch algorithm for clinical pathway clustering. J Biomed Inform. 2021;115:103668. https://doi.org/10.1016/j.jbi.2020.103668.
24. Damerau FJ. A technique for computer detection and correction of spelling errors. Commun ACM. 1964;7(3):171–6. https://doi.org/10.1145/363958.363994.
25. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Stat Assoc. 1989;84(406):414–20. https://doi.org/10.1080/01621459.1989.10478785.
26. Qu G, Yan Z, Wu H. Clover: tree structure-based efficient DNA clustering for DNA-based data storage. Brief Bioinform. 2022;23(5):bbac336. https://doi.org/10.1093/bib/bbac336.
27. Jiang M, Hu L, Han X, Zhou Y, He Z. A randomized algorithm for clustering discrete sequences. Pattern Recogn. 2024;151:110388. https://doi.org/10.1016/j.patcog.2024.110388.
28. Enright AJ. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30(7):1575–84. https://doi.org/10.1093/nar/30.7.1575.
29. Moreau C, Devogele T, Peralta V, Etienne L. A contextual edit distance for semantic trajectories. In: Proceedings of the 35th annual ACM symposium on applied computing; 2020. pp. 635–637.
30. Li R, He X, Dai C, Zhu H, Lang X, Chen W, et al. Gclust: a parallel clustering tool for microbial genomic data. Genomics Proteomics Bioinforma. 2019;17(5):496–502.
31. Bruneau M, Mottet T, Moulin S, Kerbiriou M, Chouly F, Chretien S, et al. A clustering package for nucleotide sequences using Laplacian Eigenmaps and Gaussian Mixture Model. Comput Biol Med. 2018;93:66–74.

32. Pipenbacher P, Schliep A, Schneckener S, Schönhuth A, Schomburg D, Schrader R. ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics. 2002;18(suppl_2):S182–91.
33. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
34. Belcaid M, Arisdakessian C, Kravchenko Y. Efficient DNA sequence partitioning using probabilistic subsets and hypergraphs. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing. 2021. pp. 4–9.
35. Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. 1990.
36. Halpin B. Optimal matching analysis and life-course data: the importance of duration. Sociol Methods Res. 2010;38(3):365–88. https://doi.org/10.1177/0049124110363590.
37. Namiki Y, Ishida T, Akiyama Y. Fast DNA Sequence Clustering Based on Longest Common Subsequence. In: Huang DS, Gupta P, Zhang X, Premaratne P, editors. Emerging Intelligent Computing Technology and Applications. Springer, Berlin Heidelberg: Berlin, Heidelberg; 2012. pp. 453–60.
38. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018. https://doi.org/10.1038/s41467-018-04964-5.
39. National Center for Biotechnology Information. BLAST Command Line Applications User Manual. 2008. https://www.ncbi.nlm.nih.gov/books/NBK279690/. Accessed : November 2024.
40. Emery K, Berchtold A. Comparison of two approaches in multichannel sequence analysis using the Swiss Household Panel. Longit Life Course Stud. 2023;14(4):592–623. https://doi.org/10.1332/175795921x16698302233894.
41. Gauthier JA, Widmer ED, Bucher P, Notredame C. Multichannel sequence analysis applied to social science data. Sociol Methodol. 2010;40(1):1–38.
42. Robette N, Bry X, Lelièvre v. A "Global Interdependence" approach to multidimensional sequence analysis. Sociol Methodol. 2015;45(1):1–44. https://doi.org/10.1177/0081175015570976.
43. Biemann T. A transition-oriented approach to optimal matching. Sociol Methodol. 2011;41(1):195–221.
44. Jahanshahi H, Baydogan MG. nTreeClus: a tree-based sequence encoder for clustering categorical series. Neurocomputing. 2022;494:224–41. https://doi.org/10.1016/j.neucom.2022.04.076.
45. Wei D, Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics. 2012. https://doi.org/10.1186/1471-2105-13-174.
46. Dong J, Yang X, Jiang M, Hu L, He Z. Interpretable sequence clustering. Inf Sci. 2025;689:121453. https://doi.org/10.1016/j.ins.2024.121453.
47. Millan Arias P, Hill KA, Kari L. iDeLUCS: a deep learning interactive tool for alignment-free clustering of DNA sequences. Bioinformatics. 2023;39(9):btad508. https://doi.org/10.1093/bioinformatics/btad508.
48. Nguyen D, Luo W, Nguyen TD, Venkatesh S, Phung D. Sqn2Vec: Learning Sequence Representation via Sequential Patterns with a Gap Constraint. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G, editors. Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing; 2019. pp. 569–84.
49. Millán Arias P, Alipour F, Hill KA, Kari L. DeLUCS: deep learning for unsupervised clustering of DNA sequences. PLoS One. 2022;17(1):e0261531. https://doi.org/10.1371/journal.pone.0261531.
50. Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, Salido-Ruiz RA, Vélez-Pérez H, Morales JA. Genomic signal processing for DNA sequence clustering. PeerJ. 2018;6:e4264.
51. Bruce SA. Adaptive clustering and feature selection for categorical time series using interpretable frequency-domain features. Stat Interface. 2023;16(2):319.
52. Girgis HZ. MeShClust v3.0: high-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. BMC Genomics. 2022. https://doi.org/10.1186/s12864-022-08619-0.
53. Elzinga CH, Studer M. Spell sequences, state proximities, and distance metrics. Sociol Methods Res. 2014;44(1):3–47. https://doi.org/10.1177/0049124114540707.
54. James BT, Luczak BB, Girgis HZ. MeShClust: an intelligent tool for clustering DNA sequences. Nucleic Acids Res. 2018;46(14):e83–e83. https://doi.org/10.1093/nar/gky315.
55. Peltier C, Visalli M, Schlich P, Cardot H. Analyzing temporal dominance of sensations data with categorical functional data analysis. Food Qual Prefer. 2023;109:104893. https://doi.org/10.1016/j.foodqual.2023.104893.
56. Preda C, Grimonprez Q, Vandewalle V. Categorical functional data analysis. The cfda R package. Mathematics. 2021. https://doi.org/10.3390/math9233074.
57. López-Oriona Á, Vilar JA. Analyzing categorical time series with the R package ctsfeatures. J Comput Sci. 2024;76:102233. https://doi.org/10.1016/j.jocs.2024.102233.
58. Kania A, Sarapata K. Multifarious aspects of the chaos game representation and its applications in biological sequence analysis. Comput Biol Med. 2022;151:106243. https://doi.org/10.1016/j.compbiomed.2022.106243.
59. Bollenrücher M, Darwiche J, Antonietti JP. Methodology for identification, visualization, and clustering of similar behaviors in dyadic sequences analyzed through the longitudinal actor-partner interdependence model with Markov chains. Quant Methods Psychol. 2024. https://doi.org/10.20982/tqmp.20.1.p017.
60. Tandon S, Contris D. UserJourney2Vector: Enterprise Application of Transformer Models on User Telemetry Data. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. CHI '24. ACM; 2024. pp. 1–7. https://doi.org/10.1145/3613905.3637134.
61. Frühwirth-Schnatter S, Pamminger C. Model-based clustering of categorical time series. Bayesian Anal. 2010;5(2). https://doi.org/10.1214/10-ba606.
62. Frü hwirth Schnatter S, Pamminger C, Winter-Ebmer R, Weber A, Simos TE, Psihoyios G, et al. Model-based clustering of categorical time series with multinomial logit classification. In: AIP Conference Proceedings, vol. 1281; 2010. pp. 1897.
63. Melnykov V. Model-based biclustering of clickstream data. Comput Stat Data Anal. 2016;93:31–45.
64. Melnykov V. ClickClust: an R package for model-based clustering of categorical sequences. J Stat Softw. 2016;74(9):1–34. https://doi.org/10.18637/jss.v074.i09.
65. Budel G, Frasincar F, Boekestijn D. DBHC: discrete Bayesian HMM clustering. Int J Mach Learn Cybern. 2024;15(8):3439–54. https://doi.org/10.1007/s13042-024-02102-w.
66. Murphy K, Murphy TB, Piccarreta R, Gormley IC. Clustering longitudinal life-course sequences using mixtures of exponential-distance models. J R Stat Soc Ser A Stat Soc. 2021;184(4):1414–51. https://doi.org/10.1111/rssa.12712.
67. Barban N, Billari FC. Classifying life course trajectories: a comparison of latent class and sequence analysis. J R Stat Soc C Appl Stat. 2012;61(5):765–84.
68. Linzer DA, Lewis JB. poLCA: an R package for polytomous variable latent class analysis. J Stat Softw. 2011;42(10):1–29. https://doi.org/10.18637/jss.v042.i10.
69. Zhuzhel V, Grabar V, Kaploukhaya N, Rivera-Castro R, Mironova L, Zaytsev A, et al. No Two Users Are Alike: Generating Audiences with Neural Clustering for Temporal Point Processes. Dokl Math. 2023;108(S2):S511–28. https://doi.org/10.1134/s1064562423701661.
70. Elsenburg LK, Rieckmann A, Bengtsson J, Jensen AK, Rod NH. Application of life course trajectory methods to public health data: A comparison of sequence analysis and group-based multi-trajectory modeling for modelling childhood adversity trajectories. Soc Sci Med. 2024;340:116449. https://doi.org/10.1016/j.socscimed.2023.116449.
71. Nagin DS. Group-based trajectory modeling: An overview. Ann Nutr Metab. 2014;65(2–3):205–10. https://doi.org/10.1159/000360229.
72. Gay D, Guigourès R, Boullé M, Clérot F. TESS: temporal event sequence summarization. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE; 2015. pp. 1–10.
73. Xu H, Zha H. A dirichlet mixture model of hawkes processes for event sequence clustering. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems, vol. 30. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/dd8eb9f23fbd362da0e3f4e70b878c16-Paper.pdf. Accessed : November 2024.
74. Helske S, Helske J. Mixture hidden Markov models for sequence data: the seqHMM package in R. J Stat Softw. 2019;88(3):1–32. https://doi.org/10.18637/jss.v088.i03.
75. Lu Z, Chandra NK. A sparse factor model for clustering high-dimensional longitudinal data. Stat Med. 2024;43(19):3633–48. https://doi.org/10.1002/sim.10151.
76. Vávra J, Komárek A. Classification based on multivariate mixed type longitudinal data with an application to the EU-SILC database. Adv Data Anal Classif. 2023;17(2):369–406.
77. Lin H, Han L, Peduzzi PN, Murphy TE, Gill TM, Allore HG. A dynamic trajectory class model for intensive longitudinal categorical outcome. Stat Med. 2014;33(15):2645–64.
78. Larmarange J, Briatte F, chl, Barnier J, mgdondon, Soetewey A, et al.. larmarange/analyse-R: Version du 12 octobre 2024. Zenodo; 2024. https://doi.org/10.5281/ZENODO.13923691.

79. Guyet T, Pinson P, Gesny E. Clustering of timed sequences-application to the analysis of care pathways. Data Knowl Eng. 2025;156:102401.
80. Akay Ö, Yüksel G. Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms. Commun Stat Simul Comput. 2018;47(10):3031–41.
81. Kim S, Lee J. BAG: a graph theoretic sequence clustering algorithm. Int J Data Min Bioinform. 2006;1(2):178–200.
82. Guo G, Chen L, Ye Y, Jiang Q. Cluster validation method for determining the number of clusters in categorical sequences. IEEE Trans Neural Netw Learn Syst. 2017;28(12):2936–48. https://doi.org/10.1109/tnnls.2016.2608354.
83. Hirano S, Tsumoto S. Mining Clinical Pathway Candidates from Order History Based on the Clustering of Order Sequences. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics. IEEE; 2013. pp. 3425–3429. https://doi.org/10.1109/smc.2013.584.
84. Chen J, Sun L, Guo C, Wei W, Xie Y. A data-driven framework of typical treatment process extraction and evaluation. J Biomed Inform. 2018;83:178–95. https://doi.org/10.1016/j.jbi.2018.06.004.
85. Najjar A, Gagne C, Reinharz D. Two-Step Heterogeneous Finite Mixture Model Clustering for Mining Healthcare Databases. In: 2015 IEEE International Conference on Data Mining. IEEE; 2015. pp. 931–6. https://doi.org/10.1109/icdm.2015.70.
86. Ranjan C, Paynabar K, Helm JE, Pan J. The impact of estimation: a new method for clustering and trajectory estimation in patient flow modeling. Prod Oper Manag. 2017;26(10):1893–914. https://doi.org/10.1111/poms.12722.
87. Zhang Y, Melnykov V, Zhu X. Model-based clustering of time-dependent categorical sequences with application to the analysis of major life event patterns. Stat Anal Data Min ASA Data Sci J. 2021;14(3):230–40. https://doi.org/10.1002/sam.11502.
88. Vávra J, Komárek A, Grün B, Malsiner-Walli G. Clusterwise multivariate regression of mixed-type panel data. Stat Comput. 2023;34(1). https://doi.org/10.1007/s11222-023-10304-5.
89. Jääskinen V, Parkkinen V, Cheng L, Corander J. Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. Stat Appl Genet Mol Biol. 2014;13(1). https://doi.org/10.1515/sagmb-2013-0031.
90. Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and visualizing state sequences in R with TraMineR. J Stat Softw. 2011;40:1–37.
91. Vinga S, Almeida J. Alignment-free sequence comparison–a review. Bioinformatics. 2003;19(4):513–23.
92. Fruhwirth-Schnatter S, Pamminger C, Winter-Ebmer R, Weber A, Simos TE, Psihoyios G, et al. Model-based Clustering of Categorical Time Series with Multinomial Logit Classification. In: AIP Conference Proceedings. AIP; 2010. pp. 1897–1900. https://doi.org/10.1063/1.3498286.
93. Emery K, Studer M, Berchtold A. Comparison of imputation methods for univariate categorical longitudinal data. Qual Quant. 2025;59(2):1767–91.
94. Agresti A. Categorical data analysis, vol. 792. John Wiley & Sons; 2012.
95. Yin C, Chen Y, Yau SST. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. J Theor Biol. 2014;359:18–28. https://doi.org/10.1016/j.jtbi.2014.05.043.
96. Zou Q, Lin G, Jiang X, Liu X, Zeng X. Sequence clustering in bioinformatics: an empirical study. Brief Bioinform. 2018;21(1):1–10. https://doi.org/10.1093/bib/bby090.
97. Canhão H, Faustino A, Martins F, Fonseca J. Reuma. PT - the rheumatic diseases Portuguese register. Acta Reumatol Port. 2011;36:45–56.
98. Yuan L, Hong Z, Chen L, Cai Q. Clustering Categorical Sequences with Variable-Length Tuples Representation. In: Lehner F, Fteimi N, editors. Knowledge Science, Engineering and Management. Cham: Springer International Publishing; 2016. pp. 15–27.
99. Ghassempour S, Girosi F, Maeder A. Clustering multivariate time series using hidden Markov models. Int J Environ Res Public Health. 2014;11(3):2741–63.
100. Piccarreta R, Billari FC. Clustering work and family trajectories by using a divisive algorithm. J R Stat Soc Ser A Stat Soc. 2007;170(4):1061–78. https://doi.org/10.1111/j.1467-985X.2007.00495.x.
101. Chen R, Zhang J, Ravishanker N, Konduri K. Clustering activity-travel behavior time series using topological data analysis. J Big Data Anal Transp. 2019;1(2–3):109–21. https://doi.org/10.1007/s42421-019-00008-6.
102. Borodovsky M, Peresetsky A. Deriving non-homogeneous DNA Markov chain models by cluster analysis algorithm minimizing multiple alignment entropy. Comput Chem. 1994;18(3):259–67. https://doi.org/10.1016/0097-8485(94)85022-4.
103. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53.
104. García-Magariños M, Vilar JA. A framework for dissimilarity-based partitioning clustering of categorical time series. Data Min Knowl Disc. 2015;29(2):466–502.
105. Jiang M, Wang J, Hu L, He Z. Random forest clustering for discrete sequences. Pattern Recogn Lett. 2023;174:145–51.
106. El-Yaniv R, Fine S, Tishby N. Agnostic classification of Markovian sequences. Adv Neural Inf Process Syst. 1997;10.465–71.
107. Dinu LP, Ionescu RT. Clustering based on median and closest string via rank distance with applications on DNA. Neural Comput Appl. 2014;24:77–84.
108. Saw AK, Raj G, Das M, Talukdar NC, Tripathy BC, Nandi S. Alignment-free method for DNA sequence clustering using fuzzy integral similarity. Sci Rep. 2019;9(1):3753.
109. Bao JP, Yuan RY. A wavelet-based feature vector model for DNA clustering. Genet Mol Res. 2015;14(4):19163–72. https://doi.org/10.4238/2015.december.29.26.
110. Liu L, Ho Y, Yau S. Clustering DNA sequences by feature vectors. Mol Phylogenet Evol. 2006;41(1):64–9. https://doi.org/10.1016/j.ympev.2006.05.019.
111. Mujiono WI, Veritawati I, Fractal dimension approach for clustering of DNA sequences based on internucleotide distance. In: 2013 International Conference of Information and Communication Technology (ICoICT). IEEE; 2013. pp. 82–7. https://doi.org/10.1109/icoict.2013.6574554.
112. Huang Z, Dong W, Duan H, Li H. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. IEEE J Biomed Health Inform. 2014;18(1):4–14. https://doi.org/10.1109/jbhi.2013.2274281.
113. Xu K, Chen L, Wang S. A multi-view kernel clustering framework for categorical sequences. Expert Syst Appl. 2022;197:116637. https://doi.org/10.1016/j.eswa.2022.116637.
114. Bao J, Yuan R, Bao Z. An improved alignment-free model for DNA sequence similarity metric. BMC Bioinformatics. 2014. https://doi.org/10.1186/1471-2105-15-321.
115. Wu W, Yan J, Yang X, Zha H. Discovering temporal patterns for event sequence clustering via policy mixture model. IEEE Trans Knowl Data Eng. 2022;34(2):573–86. https://doi.org/10.1109/tkde.2020.2986206.
116. Crane H. Clustering from categorical data sequences. J Am Stat Assoc. 2015;110(510):810–23. https://doi.org/10.1080/01621459.2014.983521.
117. Massoni S, Olteanu M, Villa-Vialaneix N. Which dissimilarity is to be used when extracting typologies in sequence analysis? A comparative study. In: Advances in Computational Intelligence: 12th International Work-Conference on Artificial Neural Networks, IWANN 2013, Puerto de la Cruz, Tenerife, Spain, June 12-14, 2013, Proceedings, Part I 12. Springer; 2013. pp. 69–79.
118. Mahfouz MA, Ismail MA. Efficient soft relational clustering based on randomized search applied to selection of bio-basis for amino acid sequence analysis. In: 2012 Seventh International Conference on Computer Engineering & Systems (ICCES). IEEE; 2012. pp. 287–292.
119. Maji P, Pal SK. Protein sequence analysis using relational soft clustering algorithms. Int J Comput Math. 2007;84(5):599–617.
120. Morzy T, Wojciechowski M, Zakrzewicz M. Pattern-oriented hierarchical clustering. In: East European Conference on Advances in Databases and Information Systems. Springer; 1999. pp. 179–190.
121. Le M, Nauck D, Gabrys B, Martin T. Sequential clustering for event sequences and its impact on next process step prediction. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems: 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part I 15. Springer; 2014. pp. 168–178.
122. Dehghanzadeh H, Ghaderi-Zefrehei M, Mirhoseini SZ, Esmaeilkhaniyan S, Haruna IL, Amirpour NH. Retracted article: a new DNA sequence entropy-based Kullback-Leibler algorithm for gene clustering. J Appl Genet. 2020;61(2):231–8. https://doi.org/10.1007/s13353-020-00543-x.
123. Aleb N, Labidi N, An improved K-means algorithm for DNA sequence clustering. In: 2015 26th International Workshop on Database and Expert Systems Applications (DEXA). IEEE; 2015. pp. 39–42.
124. Oh SJ, Kim JY. A scalable clustering method for categorical sequence data. Int J Comput Methods. 2005;2(02):167–80.
125. Mao Q, Zheng W, Wang L, Cai Y, Mai V, Sun Y. Parallel hierarchical clustering in linearithmic time for large-scale sequence analysis. In: 2015 IEEE International Conference on Data Mining. IEEE; 2015. pp. 310–319.
126. Pham TD, Zuegg J. A probabilistic measure for alignment-free sequence comparison. Bioinformatics. 2004;20(18):3455–61.

127. Chikhaoui B, Wang S, Xiong T, Pigot H. Pattern-based causal relationships discovery from event sequences for modeling behavioral user profile in ubiquitous environments. Inf Sci. 2014;285:204–22.
128. Funkner AA, Yakovlev AN, Kovalchuk SV. Towards evolutionary discovery of typical clinical pathways in electronic health records. Procedia Comput Sci. 2017;119:234–44.
129. Oh SJ, Kim JY. A hierarchical clustering algorithm for categorical sequence data. Inf Process Lett. 2004;91(3):135–40.
130. Akay Ö, Yüksel G. Hierarchical clustering of mixed variable panel data based on new distance. Commun Stat-Simul Comput. 2021;50(6):1695–710.
131. Giannoula A, Comas M, Castells X, Estupiñán-Romero F, Bernal-Delgado E, Sanz F, et al. Exploring long-term breast cancer survivors' care trajectories using dynamic time warping-based unsupervised clustering. J Am Med Inform Assoc. 2024;31(4):820–31.
132. Ritschard G, Liao TF, Struffolino E. Strategies for multidomain sequence analysis in social research. Sociol Methodol. 2023;53(2):288–322.
133. Piccarreta R, Struffolino E. Tools for analysing fuzzy clusters of sequences data. Demogr Res. 2024;51:553–76.
134. Egho E, Raïssi C, Calders T, Jay N, Napoli A. On measuring similarity for sequences of itemsets. Data Min Knowl Disc. 2015;29:732–64.
135. Haghir Ebrahim-Abadi M, Fatemizadeh E. A Clustering-Based Algorithm for De Novo Motif Discovery in DNA Sequences. In: 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME). IEEE; 2017. pp. 1–6. https://doi.org/10.1109/icbme.2017.8430242.
136. Guralnik V, Karypis G. A scalable algorithm for clustering sequential data. In: Proceedings 2001 IEEE International Conference on Data Mining. ICDM-01. IEEE Comput. Soc; 2001. pp. 179–186. https://doi.org/10.1109/icdm.2001.989516.
137. Huang HH, Yu C. Clustering DNA sequences using the out-of-place measure with reduced n-grams. J Theor Biol. 2016;406:61–72. https://doi.org/10.1016/j.jtbi.2016.06.029.
138. Magallanes J, Stone T, Morris PD, Mason S, Wood S, Villa-Uriol MC. Sequen-C: a multilevel overview of temporal event sequences. IEEE Trans Visual Comput Graphics. 2022;28(1):901–11. https://doi.org/10.1109/tvcg.2021.3114868.
139. Yuan L, Wang W, Chen L. Two-stage pruning method for gram-based categorical sequence clustering. Int J Mach Learn Cybern. 2017;10(4):631–40. https://doi.org/10.1007/s13042-017-0744-y.
140. Chen R, Sun H, Chen L, Zhang J, Wang S. Dynamic order Markov model for categorical sequence clustering. J Big Data. 2021. https://doi.org/10.1186/s40537-021-00547-2.
141. Xiong T, Wang S, Jiang Q, Huang JZ. A New Markov Model for Clustering Categorical Sequences. In: 2011 IEEE 11th International Conference on Data Mining. IEEE; 2011. pp. 854–863. https://doi.org/10.1109/icdm.2011.13.
142. Volkovich Z, Kirzhner V, Bolshoy A, Nevo E, Korol A. The method of -grams in large-scale clustering of DNA texts. Pattern Recogn. 2005;38(11):1902–12. https://doi.org/10.1016/j.patcog.2005.05.002.
143. Ramírez V, Román-Godínez I, Torres-Ramos S. In: DNA-MC: Tool for Mapping and Clustering DNA Sequences. Springer International Publishing; 2019. pp. 736–742. https://doi.org/10.1007/978-3-030-30648-9_98.
144. Dakhli A, Bellil W, Ben Amar C. In: DNA Sequence Classification Using Power Spectrum and Wavelet Neural Network. Springer International Publishing; 2017. pp. 391–402. https://doi.org/10.1007/978-3-319-52941-7_39.
145. Yin C, Yin XE, Wang J. A novel method for comparative analysis of DNA sequences by Ramanujan-Fourier Transform. J Comput Biol. 2014;21(12):867–79. https://doi.org/10.1089/cmb.2014.0120.
146. Sangma JW, Yogita, Pal V, Kumar N, Kushwaha R. FHC-NDS: Fuzzy Hierarchical Clustering of Multiple Nominal Data Streams. IEEE Trans Fuzzy Syst. 2023;31(3):786–98. https://doi.org/10.1109/tfuzz.2022.3189083.
147. Vrotsou K, Ynnerman A, Cooper M. Are we what we do? Exploring group behaviour through user-defined event-sequence similarity. Inf Vis. 2013;13(3):232–47. https://doi.org/10.1177/1473871613477852.
148. De Angelis L, Dias JG. Mining categorical sequences from data using a hybrid clustering method. Eur J Oper Res. 2014;234(3):720–30. https://doi.org/10.1016/j.ejor.2013.11.002.
149. Zhang J, Konduri K, Chen R, Ravishanker N. Novel divide and combine-based approach to estimating mixture Markov model for large categorical time series data: application to study of clusters using multiyear travel survey data. Transp Res Rec J Transp Res Board. 2019;2673:036119811984747. https://doi.org/10.1177/0361198119847476.
150. Zhu X. Mixture modelling of categorical sequences with secondary components. Stat. 2020. https://doi.org/10.1002/sta4.295.
151. Frühwirth-Schnatter S, Pittner S, Weber A, Winter-Ebmer R. Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering. Ann Appl Stat. 2018;12(3). https://doi.org/10.1214/17-aoas1132.
152. Urso F, Abbruzzo A, Chiodi M, Cracolici MF. Model selection for mixture hidden Markov models: an application to clickstream data. Stat Pap. 2024;65(9):5797–834. https://doi.org/10.1007/s00362-024-01608-3.

Publisher's Note