Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 111
DOI: 10.11159/icsta22.111

Unsupervised Classification of Categorical Time Series through Innovative Distances

Ángel López-Oriona[1], José A. Vilar[1], Pierpaolo D'Urso[2]
[1]Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, Spain
oriona38@hotmail.com; jose.vilarf@udc.es
[2]Department of Social Sciences and Economics, Sapienza University of Rome, Italy
pierpaolo.durso@uniroma1.it

Abstract **-** In this paper, two novel distances for nominal time series are introduced. Both of them are based on features describing the serial dependence patterns between each pair of categories. The first dissimilarity employs the so-called association measures, whereas the second computes correlation quantities between indicator processes whose uniqueness is guaranteed from standard stationary conditions. The metrics are used to construct crisp algorithms for clustering categorical series. The approaches are able to group series generated from similar underlying stochastic processes, achieve accurate results with series coming from a broad range of models and are computationally efficient. An extensive simulation study shows that the devised clustering algorithms outperform several alternative procedures proposed in the literature. Specifically, they achieve better results than approaches based on maximum likelihhod estimation, which take advantage of knowing the real underlying procedures. Both innovative dissimilarities could be useful for practitioners in the field of time series clustering.

Keywords**:** categorical time series, clustering, association measures, indicator processes

1. Introduction

Clustering of time series concerns the challenge of splitting a set of unlabeled time series into homogeneous groups, which is a pivotal problem in many knowledge discovery tasks [1]. Categorical time series (CTS) are a particular class of time series exhibiting a qualitative range which consists of a finite number of categories. Most of the classical statistical tools used for real-valued time series (e.g., the autocorrelation function) are not useful in the categorical case, so different types of measures than the standard ones are needed for a proper analysis of CTS. CTS arise in an extensive assortment of fields [2, 3]. Since only a few works have addressed the problem of CTS clustering [4, 5], the main goal of this paper is to introduce novel clustering algorithms for CTS.

2. Two Novel Feature-Based Approaches for Categorical Time Series Clustering

Consider a set of S categorical time series $\mathcal{S} = \{X_t^{(1)}, \ldots, X_t^{(s)}\}$, where the j-th element is a T_j length partial realization from any categorical stochastic process $(X_t)_{t\in\mathbb{Z}}$. We suppose that the process $(X_t)_{t\in\mathbb{Z}}$ is is bivariate stationary, i.e., the pairwise joint distribution of (X_{t-k}, X_t) is invariant in t. Additionally, it is assumed that the range of the process is coded as $\mathcal{V} = \{1, \ldots, r\}$. Our goal is to perform clustering on the elements of $\mathcal{S}$ in such a way that the series generated from identical stochastic processes are placed together. To that aim, we propose two distance metrics which are based on feature extraction.

2.1. Descriptive features for categorical processes

Features based on association measures

Let $\{X_t, t \in \mathbb{Z}\}$ be a bivariate stationary categorical stochastic process with range $\mathcal{V} = \{1, \dots, r\}$. Denote by $\boldsymbol{\pi} = (\pi_1, \dots, \pi_r)$ the marginal distribution of X_t, which is, $P(X_t = j) = \pi_j > 0$, $j = 1, \dots, r$. Fixed $l \in \mathbb{N}$ and $i, j \in \mathcal{V}$, we define the lagged bivariate probability $p_{ij}(l)$ as $p_{ij}(l) = P(X_t = i, X_{t-l} = j)$.

Note that the quantity $p_{ij}(l)$ measures the serial independence at lag l between the categories j and i. In fact, this quantity is used to define the concept of perfect serial independence at lag $l \in \mathbb{N}$, in the sense that such independence exists if $p_{ij}(l) = \pi_i \pi_j$ for any $i, j \in \mathcal{V}$. There are several association measures which describe the serial dependence structure of a categorical process at lag l. One of such measures is the so-called Cramer's v, which is defined as

$$v(l) = \sqrt{\frac{1}{r-1} \sum_{i,j=1}^{r} \frac{(p_{ij}(l) - \pi_i \pi_j)^2}{\pi_i \pi_j}}. \quad (1)$$

Cramer's v summarizes the serial dependence patterns of a categorical process for every pair (i, j) and $l \in \mathbb{N}$. However, this quantity is not appropriate for describing a given stochastic process, since two different processes can have the same value of $v(l)$. A better way to characterize the process X_t is by considering the matrix $\boldsymbol{V}(l) = (V_{ij}^l)_{1 \le i,j \le r}$, where $V_{ij}^l = \frac{(p_{ij}(l) - \pi_i \pi_j)^2}{\pi_i \pi_j}$. The quantities in the matrix $\boldsymbol{V}(l)$ give information about the so-called *unsigned* dependence of the process. However, it is often useful to know whether a process tends to stay in the state it has reached or, on the contrary, the repetition of the same state after l steps is infrequent. This motivates the concept of *signed* dependence, which arises as an analogy of the autocorrelation function of a numerical process, since such quantity can take either positive or negative values. Let's first define the conditional bivariate probabilities at lag l as $p_{i|j} = P(X_t = i | X_{t-l} = j) = \frac{p_{ij}(l)}{\pi_j}$ for $i, j = 1, \dots, r$.

According to the previous quantities, perfect serial dependence occurs if, for any $j \in \mathcal{V}$, the conditional distribution of $p_{\cdot|j}(l)$ is a one-point distribution. Provided that perfect serial dependence holds, we have perfect *positive* (*negative*) serial dependence if $p_{i|i}(l) = 1$ ($p_{i|i}(l) = 0$) for all $i \in \mathcal{V}$.

A common measure of signed serial dependence at lag l is the Cohen's κ

$$\kappa(l) = \frac{\sum_{j=1}^{r} (p_{jj}(l) - \pi_j^2)}{1 - \sum_{j=1}^{r} \pi_j^2}. \quad (2)$$

Proceeding as with $v(l)$, the quantity $\kappa(l)$ can be decomposed in order to obtain a complete representation of the signed dependence patterns of the process. In this way, we consider the vector $\boldsymbol{\mathcal{K}}(l) = (\mathcal{K}_1(l), \dots, \mathcal{K}_r(l))$, where each $\mathcal{K}_i$ is defined as

$$\mathcal{K}_i(l) = \frac{p_{ii}(l) - \pi_i^2}{1 - \sum_{j=1}^{r} \pi_j^2}, \quad (3)$$

$i = 1, \dots, r$.

In practice, the matrix $\boldsymbol{V}(l)$ and the vector $\boldsymbol{\mathcal{K}}(l)$ must be estimated from a T-length realization of the process, $\{X_1, \dots X_T\}$. To this aim, we consider estimators of π_i and $p_{ij}(l)$, π_i and $p_{ij}(l)$, respectively, defined as$\pi_i = \frac{N_i}{T}$ $and p_{ij}(l) = \frac{N_{ij}(l)}{T-l}$, where N_i is the number of variables X_t equal to i in the realization $\{X_1, \dots X_T\}$, and $N_{ij}(l)$ is the number of pairs $(X_t, X_{t-l}) = (i, j)$ in the realization $\{X_1, \dots X_T\}$. Hence, estimates of $\boldsymbol{V}(l)$ and $\boldsymbol{\mathcal{K}}(l)$, $\boldsymbol{V}(l)$ and $\boldsymbol{\mathcal{K}}(l)$, respectively, can be obtained by plugging in the estimates π_i and $p_{ij}(l)$ in (2) and (3). This leads directly to estimates of $v(l)$ and $\kappa(l)$, so-called $v(l)$ and $\kappa(l)$.

Properties of Cramer's v and Cohen's κ are given in [7]. In particular, Cramer's v has range $[0,1]$, with the values 0 and 1 associated with the cases of perfect serial independence and perfect serial dependence at lag l, respectively. On the other hand, The range of Cohen's κ is given by $[-\sum_{j=1}^{r} \pi_j^2/(1 - \sum_{j=1}^{r} \pi_j^2), 1]$, with the lower and upper bounds associated with the cases of perfect negative and perfect positive dependence, respectively.

Features Based On Indicator Processes

An alternative way of describing the dependence structure of the process $\{X_t, t \in \mathbb{Z}\}$ is by defining auxiliary processes showing the occurrence of each category. Given $i \in \mathcal{V}$, consider the process $I_t^i = I(X_t = i)$, where *I* stands for the indicator function. Fixed $l \in \mathbb{N}$ and $i, j \in \mathcal{V}$, consider the correlation

$$\phi_{ij}(l) = Corr(I_t^i, I_{t-l}^j), \tag{4}$$

which measures linear dependence between the indicator process of the *j*-th category and the indicator process of the i-th category l instants later. The following Lemma provides some properties of the quantity $\phi_{ij}(l)$.

Lemma 1.
Let $\{X_t, t \in \mathbb{Z}\}$ be a bivariate stationary categorical process with range $\mathcal{V} = \{1, \dots, r\}$. Then the following properties hold:

1. For every $i, j \in \mathcal{V}$, the function $\phi_{ij}: \mathbb{N} \to [-1,1]$ given by $l \to \phi_{ij}(l) = Corr(I_t^i, I_{t-l}^j)$ is well-defined.
2. $\phi_{ij}(l) = 0 \Leftrightarrow p_{ij}(l) = \pi_i \pi_j$.
3. $\phi_{ij}(l) = \pm 1 \Leftrightarrow p_{ij}(l) = \pm\sqrt{\pi_i(1-\pi_i)\pi_j(1-\pi_j)} + \pi_i \pi_j$.
4. $\phi_{ij}(l) = \sqrt{\frac{\pi_j(1-\pi_i)}{\pi_i(1-\pi_j)}} \Leftrightarrow p_{i|j}(l) = 1$.

The proof of Lemma 1 is quite straightforward and it is not shown in the manuscript for the sake of brevity. According to Lemma 1, the quantity $\phi_{ij}(l)$ can be used to explain both types of dependence, signed and unsigned, within the underlying process. In fact, in the case of perfect unsigned independence at lag l, we have that $p_{ij}(l) = \pi_i \pi_j$ for all

$i, j \in \mathcal{V}$ so that $\phi_{ij}(l) = 0$ for all $i, j \in \mathcal{V}$ in accordance with Property 2. On the other hand, when we have perfect positive dependence at lag l, then $p_{i|i}(l) = 1$ for all $i \in \mathcal{V}$. Then $\phi_{ii}(l) = 1$ for all $i \in \mathcal{V}$ by following Property 4. Therefore, $\phi_{ij}(l)$ evaluates unsigned dependence when $i \neq j$ and signed dependence when $i = j$. The quantities in (4) can be encapsulated in a matrix $\boldsymbol{\Phi}(l) = (\phi_{ij}(l))_{1 \leq i,j \leq r}$. The matrix $\boldsymbol{\Phi}(l)$ can be easily estimated by means of $\boldsymbol{\Phi}(l) = (\phi_{ij}(l))_{1 \leq i,j \leq r}$, where each estimator $\phi_{ij}(l)$ is computed as $\phi_{ij}(l) = \frac{p_{ij}(l) - \pi_i \pi_j}{\sqrt{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}}$, which follows from the fact that $\phi_{ij}(l) = \frac{E(I_t^i I_{t-l}^j) - E(I_t^i)E(I_{t-l}^j)}{\sqrt{Var(I_t^i)Var(I_{t-l}^j)}}$.

2.2. Two innovative dissimilarities between CTS

In this section we introduce two distance measure between categorical series based on the features described above. Suppose we have a pair of CTS $X_t^{(1)}$ and $X_t^{(2)}$ and consider a set of L lags, $\mathcal{L} = \{l_1, \dots, l_L\}$. A dissimilarity based on association measures, so-called d_{AM}, is defined as

$$d_{AM}(X_t^{(1)}, X_t^{(2)}) = \sum_{k=1}^{L} [||vec(\boldsymbol{V}(l_k)^{(1)} - \boldsymbol{V}(l_k)^{(2)})||^2 + ||\boldsymbol{\mathcal{K}}(l_k)^{(1)} - \boldsymbol{\mathcal{K}}(l_k)^{(2)}||^2] + ||\boldsymbol{\pi}^{(1)} - \boldsymbol{\pi}^{(2)}||^2 =$$
$$\sum_{k=1}^{L} \sum_{i=1}^{r} \sum_{j=1}^{r} (V_{ij}^{l_k(1)} - V_{ij}^{l_k(2)})^2 + \sum_{k=1}^{L} \sum_{i=1}^{r} (\mathcal{K}_i(l_k)^{(1)} - \mathcal{K}_i(l_k)^{(2)}) + \sum_{i=1}^{r} (\pi_i^{(1)} - \pi_i^{(2)})^2, \qquad (5)$$

where the superscripts (1) and (2) are used to indicate that the corresponding estimation is obtained with respect to the realization $X_t^{(1)}$ and $X_t^{(2)}$, respectively.

An alternative distance measure relying on indicator processes, so-called d_{IP}, is defined as

$$d_{IP}(X_t^{(1)}, X_t^{(2)}) = \sum_{k=1}^{L} [||vec(\boldsymbol{\Phi}(l_k)^{(1)} - \boldsymbol{\Phi}(l_k)^{(2)})||^2] + ||\boldsymbol{\pi}^{(1)} - \boldsymbol{\pi}^{(2)}||^2 =$$
$$\sum_{k=1}^{L} \sum_{i=1}^{r} \sum_{j=1}^{r} (\phi_{ij}(l_k)^{(1)} - \phi_{ij}(l_k)^{(2)})^2 + \sum_{i=1}^{r} (\pi_i^{(1)} - \pi_i^{(2)})^2. \qquad (6)$$

For a given set of categorical series, the distances d_{AM} and d_{IP} can be used as input for traditional clustering algorithms. In this manuscript we consider the *Partition Around Medoids* (PAM) algorithm.

3. Partitioning Around Medoids Clustering Of Categorical Time Series

In this section we examine the performance of both metrics d_{AM} and d_{IP} in the context of hard clustering through a simulation study.

3.1 Experimental design

The simulated scenarios encompass a broad variety of generating processes. In particular, three setups were considered, namely clustering of (i) Markov Chains (MC), (ii) New Discrete ARMA (NDARMA) processes and (iii) Hidden Markov Models (HMM). The generating models with respect to each class of processes are given below.

Scenario 1. Clustering of MC. Consider four three-state MC, so-called MC$_1$, MC$_2$, MC$_3$ and MC$_4$, with respective transition matrices $\boldsymbol{P}_1^1$, $\boldsymbol{P}_2^1$, $\boldsymbol{P}_3^1$ and $\boldsymbol{P}_4^1$ given by

$$\boldsymbol{P}_1^1 = Mat^3(0.1,0.8,0.1,0.5,0.4,0.1,0.6,0.2,0.2),$$
$$\boldsymbol{P}_2^1 = Mat^3(0.1,0.8,0.1,0.6,0.3,0.1,0.6,0.2,0.2),$$
$$\boldsymbol{P}_3^1 = Mat^3(0.05,0.90,0.05,0.05,0.05,0.90,0.90,0.05,0.05),$$
$$\boldsymbol{P}_4^1 = Mat^3(1/3,1/3,1/3,1/3,1/3,1/3,1/3,1/3,1/3),$$

where the operator Mat^k, $k \in \mathbb{N}$ transforms a vector into a square matrix of order k by sequentially placing the corresponding numbers by rows.

Scenario 2. Clustering of HMM. Consider the bivariate process $(X_t, Q_t)_{t\in\mathbb{Z}}$, where Q_t stands for the hidden states and $X_t$for the observable random variables. Process $(Q_t)_{t\in\mathbb{Z}}$ constitutes an homogeneous MC. Both $(X_t)_{t\in\mathbb{Z}}$ and $(Q_t)_{t\in\mathbb{Z}}$ are assumed to be count processes with range $\{1, \dots, r\}$. Process $(X_t, Q_t)_{t\in\mathbb{Z}}$ is assumed to verify the three classical assumptions of a HMM. Based on previous considerations, let HMM$_1$, HMM$_2$, HMM$_3$ and HMM$_4$ be four three-state HMM with respective transition matrices $\boldsymbol{P}_1^2$, $\boldsymbol{P}_2^2$, $\boldsymbol{P}_3^2$ and $\boldsymbol{P}_4^2$ and emission matrices $\boldsymbol{E}_1^2$, $\boldsymbol{E}_2^2$, $\boldsymbol{E}_3^2$ and $\boldsymbol{E}_4^2$ given by

$$\boldsymbol{P}_1^2 = Mat^3(0.05,0.90,0.05,0.05,0.05,0.90,0.90,0.05,0.05),\ \boldsymbol{P}_2^2 = \boldsymbol{P}_1^2,$$
$$\boldsymbol{P}_3^2 = Mat^3(0.1,0.7,0.2,0.4,0.4,0.2,0.4,0.3,0.3),$$
$$\boldsymbol{P}_4^2 = Mat^3(1/3,1/3,1/3,1/3,1/3,1/3,1/3,1/3,1/3),\ \boldsymbol{E}_1^2 = \boldsymbol{P}_1^2,$$
$$\boldsymbol{E}_2^2 = Mat^3(0.1,0.8,0.1,0.5,0.4,0.1,0.6,0.2,0.2),\ \boldsymbol{E}_3^2 = \boldsymbol{E}_2^2,$$
$$\boldsymbol{E}_4^2 = Mat^3(1/3,1/3,1/3,1/3,1/3,1/3,1/3,1/3,1/3).$$

Scenario 3. Clustering of NDARMA processes. Let $(X_t)_{t\in\mathbb{Z}}$ and $(\epsilon_t)_{t\in\mathbb{Z}}$, be two count processes with range $\{1, \dots, r\}$following the equation

$$X_t = \alpha_{t,1}X_{t-1} + \cdots + \alpha_{t,p}X_{t-p} + \beta_{t,0}\epsilon_t + \cdots + \beta_{t,q}\epsilon_{t-q},$$

where $(\epsilon_t)_{t\in\mathbb{Z}}$ is i.i.d with $P(\epsilon_t = i) = \pi_i$, independent of $(X_s)_{s<t}$, and the i.i.d multinomial random vectors

$$(\alpha_{t,1}, \dots, \alpha_{t,p}, \beta_{t,0}, \dots, \beta_{t,q}) \sim MULT(1; \phi_1, \dots, \phi_p, \varphi_0, \dots, \varphi_q),$$

are independent of $(\epsilon_t)_{t\in\mathbb{Z}}$ and $(X_s)_{s<t}$. The considered models are three three-state NDARMA(2,0) processes and one three-state NDARMA(1,0) process with marginal distribution $\boldsymbol{\pi}^3 = (2/3,1/6,1/6)$, and corresponding probabilities in the multinomial distribution given by

$(\phi_1, \phi_2, \varphi_0)_1^3 = (0.7,0.2,0.1)$, $(\phi_1, \phi_2, \varphi_0)_2^3 = (0.1,0.45,0.45)$, $(\phi_1, \phi_2, \varphi_0)_3^3 = (0.5,0.25,0.25)$, $(\phi_1, \varphi_0)_4^3 = (0.2,0.8)$.

The simulation study was carried out as follows. For each scenario, 5 CTS of length $T = 200,600$ were generated from each process in order to execute the clustering techniques twice, thus allowing to analyze the impact of the series length. The resulting experimental solution produced by the PAM algorithm was stored. The simulation procedure was repeated 500 times for each scenario and value of T. The computation of d_{AM} and d_{IP} was carried out by considering $\mathcal{L} = \{1\}$ in Scenarios 1 and 2, and $\mathcal{L} = \{1,2\}$ in Scenario 3. This way, we adapted the distances to the maximum number of significant lags existing in each scenario.

3.2 Alternative metrics and assessment criteria

To better analyze the performance of both metrics d_{AM} and d_{IP}, we also obtained partitions by using alternative techniques for clustering of categorical series. The considered procedures are described below.

- *Model-based approach using maximum likelihood estimation (MLE).* The distance between two CTS is defined as the squared Euclidean distance between the corresponding vectors of fitted coefficients via MLE (d_{MLE}).
- *Model-based approach using mixtures*. [4] propose to cluster a set of CTS by learning a mixture of first order Markov models via the EM algorithm (d_{CZ}).
- *An hybrid framework for clustering CTS*. [6] presents a dissimilarity between categorical series which evaluates both closeness between raw categorical values and proximity between dynamic patterns (d_{MV}).

Note that the approach based on the distance d_{MLE} can be seen as a strict benchmark in the evaluation task. The effectiveness of the clustering approaches was assessed by comparing the clustering solution produced by the algorithms with the true clustering partition, so-called ground truth. The latter consisted of $C = 4$ clusters in all scenarios, each group including the five CTS generated from the same process. The value $C = 4$ was provided as input parameter to the PAM algorithm in the case of d_{AM}, d_{IP}, d_{MLE} and d_{MV}. Experimental and true partitions were compared by using three well-known external clustering quality indexes, the Adjusted Rand Index (ARI), the Jaccard Index (JI) and the Fowlkes-Mallows index (FMI).

3.3 Results and discussion

Average values of the quality indexes by taking into account the 500 simulation trials are given in Tables 1, 2 and 3 for Scenarios 1, 2 and 3, respectively.

Table 1: Average results for Scenario 1.

	$T = 200$			$T = 600$		
Method	ARI	JI	FMI	ARI	JI	FMI
d_{AM}	**0.77**	**0.71**	**0.83**	**0.92**	**0.89**	0.94
d_{IP}	0.73	0.66	0.79	0.86	0.88	0.89
d_{MLE}	0.70	0.63	0.77	0.84	0.79	0.88
d_{CZ}	0.71	0.65	0.79	0.92	**0.89**	0.93
d_{MV}	0.41	0.36	0.67	0.38	0.36	0.65

Table 1: Average results for Scenario 2.

	$T = 200$			$T = 600$		
Method	ARI	JI	FMI	ARI	JI	FMI
d_{AM}	0.71	0.64	0.78	0.86	0.81	0.89
d_{IP}	**0.76**	**0.70**	**0.81**	**0.96**	**0.95**	**0.97**
d_{MLE}	0.35	0.34	0.51	0.30	0.31	0.48
d_{CZ}	0.65	0.58	0.74	0.70	0.64	0.78
d_{MV}	0.09	0.18	0.32	0.06	0.18	0.30

Table 1: Average results for Scenario 3.

	$T = 200$			$T = 600$		
Method	ARI	JI	FMI	ARI	JI	FMI
d_{AM}	0.63	0.56	0.72	0.88	0.84	0.90
d_{IP}	0.68	0.61	0.75	**0.93**	**0.90**	**0.94**
d_{MLE}	**0.73**	**0.66**	**0.79**	0.87	0.83	0.90
d_{CZ}	0.59	0.56	0.69	0.65	0.58	0.74
d_{MV}	0.04	0.17	0.29	-0.03	0.14	0.25

The results in Table 1 indicate that the dissimilarity d_{AM} is the best performing one when dealing with MC, outperforming the MLE-based metric d_{MLE}. The distance d_{IP} is also superior to d_{MLE}. The measure d_{CZ} attains in Scenario 1 similar results than d_{AM}, specially for $T = 600$. The good performance of d_{CZ} was expected, since the assumption of first order Markov models considered by this metric is fulfilled in Scenario 1. Table 2 shows a completely different picture, indicating that the metrics d_{AM} and d_{IP} exhibit a significantly better effectiveness than the rest of the dissimilarities. Finally, the quantities in Table 3 reveal that the model-based distance d_{MLE} attains the best results when $T = 200$, but is defeated by d_{IP} when $T = 600$. The metric d_{CZ} suffers again from model misspecification. In summary, the numerical experiments carried out throughout this section show the excellent ability of both measures d_{AM} and d_{IP} to discriminate between a broad variety of categorical processes. Specifically, these metrics either outperform or show similar behavior than distances based on estimated model coefficients, which take advantage of knowing the true underlying models.

4. Conclusions

In this paper we introduced two metrics to perform cluster analysis of categorical series. The goal of both distances is to discriminate between underlying categorical processes. The two dissimilarities are used to construct clustering algorithms, which were evaluated in a broad simulation study. The methods outperform several alternatives proposed in the literature, suggesting the usefulness of the proposed dissimilarities for clustering of categorical series.

References

[1] T. W. Liao, "Clustering of time series data: A survey," Pattern Recognit., vol. 38, no. 11, pp. 1857–1874, 2005.

[2] G. A. Churchill, "Stochastic models for heterogeneous dna sequences," Bulletin of mathematical biology, vol. 51, no. 1, pp. 79–94, 1989.

[3] K. Fokianos and B. Kedem, "Regression theory for categorical time series," Statistical science, vol. 18, no. 3, pp. 357–376, 2003.

[4] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Model-based clustering and visualization of navigation patterns on a web site," Data mining and knowledge discovery, vol. 7, no. 4, pp. 399–424, 2003.

[5] S. Frühwirth-Schnatter and C. Pamminger, "Model-based clustering of categorical time series," Bayesian Analysis, vol. 5, no. 2, pp. 345–368, 2010.

[6] M. García-Magariños and J. A. Vilar, "A framework for dissimilarity- based partitioning clustering of categorical time series," Data mining and knowledge discovery, vol. 29, no. 2, pp. 466–502, 2015.

[7] Weiß, C. H. (2018). An introduction to discrete-valued time series. John Wiley & Sons.