Case Study Using the Generalized Framework for Optimal Temporal Clustering

by

Jiazhou Liang

A MEng project report submitted in conformity with the requirements
for the degree of Master of Engineering

Department of Mechanical and Industrial Engineering
University of Toronto

Case Study Using the Generalized Framework for Optimal Temporal Clustering

Jiazhou Liang
Master of Engineering
Department of Mechanical and Industrial Engineering
University of Toronto
2024

Abstract

Clustering, a widely employed unsupervised machine-learning tool, has found applications spanning diverse disciplines. Temporal clustering, particularly the task of grouping unlabelled multivariate time-series data, has attracted considerable attention from researchers [1]. Jolomi Tosanwumi, a current Master of Applied Science student, has recently introduced a generalized framework for temporal clustering that effectively addresses two critical issues inherent in prior algorithms: (1) the incapacity to analyze cluster changes over time and (2) suboptimal outcomes arising from different initializations. In collaboration with Tosanwumi, this project undertook a comparative performance analysis of the newly proposed framework against existing algorithms using synthetic data. Additionally, two case studies were conducted employing real-world census and climate temporal data to validate the efficacy of the proposed temporal clustering framework. The results of the case studies unveiled potential applications through analytical insights, including the discernment of coastal and inland cities using worldwide historical climate data, the revelation of a potential migration pattern in the workforce with occupations in art and culture in the city of Toronto, and the identification of two distinct trends in the rental market of Downtown Toronto.

Contents

Chapter 1

Introduction

1.1 Background

We live in a world replete with various types of data for information encountered in daily lives or measured by researchers in their studies. The measurement of time stands out as a crucial feature, integral to nearly every scientific experiment [2]. Time series stores information about a feature in a sequence of data points ordered by time. It is increasingly utilized across various fields, including forecasting, financial analysis, and sociology studies [3]. When combining multiple time series, the resulting (multidimensional) temporal data encapsulates the evolution of one or more descriptive features of an object over time [4]. Analyzing multidimensional temporal data aids researchers in uncovering historical trends and patterns related to the study subjects [5].

Cluster analysis, also referred to as clustering, plays a pivotal role in modern data science and analytics, particularly when exploring datasets with limited prior knowledge of the samples within them [6]. Clustering algorithms, viewed from a mathematical perspective, entail learning sets of groups of samples through an unsupervised process. The objective is to achieve, within a given metric space, groups of data points that exhibit the highest similarity relative to data points in other groups [7, 4]. Temporal clustering, a specific aspect of clustering, involves clustering multiple time series data in a manner that ensures cluster assignments remain coherent across successive time points [7, 8]. Temporal clustering holds the potential to unveil valuable information and trends within temporal data [1]. Furthermore, it serves as a succinct summary of the data, enabling a quick comprehension of the overall structure and facilitating the detection of anomalies or outliers [4].

The current temporal clustering algorithm can be categorized into two main groups:

1. **Time Series Clustering (TSC)** involves clustering temporal time series by treating them as multidimensional vectors [9]. While the clusters themselves are dynamic and the clustering process considers time, cluster labels are not permitted to change.

2. **Temporal Label Analysis (TLA)** involves pre-clustering without considering time and then analyzing how cluster labels change over time [10, 11]. While cluster labels can change, the cluster definitions are not allowed to dynamically evolve over time.

Nevertheless, current temporal clustering algorithms exhibit several deficiencies. Firstly, these methods lack optimality guarantees and suffer from inconsistent reproducibility due to the sensitivity of their solutions

to (often randomized) initialization. Although some research has shown that enhanced initialization can improve the performance of underlying temporal clustering methods, such as K-means [12], the assurance of optimality remains elusive.

Secondly, as previously mentioned, neither TSC nor TLA permits dynamic changes in cluster centers and labels simultaneously. This limitation confines them to specific settings and leaves a methodological design space underexplored. This restriction hampers their adaptability and versatility in handling diverse data scenarios.

1.2 Generalized Framework for Optimal Temporal Clustering

To address these deficiencies, Jolomi Tosanwumi, a current Master of Applied Science student in the Data-Driven Decision Making Lab at the University of Toronto Department of Mechanical and Industrial Engineering has presented generalized design space for temporal clustering based on two key choices: (1) Are cluster definitions (centers) Z allowed to dynamically change over time? (2) Are time series entities (e.g., locations) allowed to change cluster assignment C over time, and if so, how much?

Leveraging this design space, he first proposed six possible cluster schemes, and together, we designed Table 1.1 to provide a summarization and descriptions of each scheme in the design space.

	Fixed Cluster Center Z	**Temporally Dynamic** Cluster Center Z
Fixed Cluster Assignment C	**Fixed Time Series Clustering (FTSC)** Cluster centers are fixed and each time series cannot change cluster over time.	**Standard Time Series Clustering (STSC)** Cluster centers can change but each time series cannot change cluster over time. (Simply referred to as Time Series Clustering by most researchers).
Unbounded Cluster Assignment C	**Fixed Unconstrained Temporal Clustering (FUTC)** Cluster centers are fixed while each time series is allowed to change clusters over time without constraints.	**Unconstrained Temporal Clustering (UTC)** When both the cluster definitions and assignments change, it results in a traditional static clustering (points in each time series are not chronological).
Cluster Assignment C with **Bounded Change**	**Fixed Constrained Temporal Clustering (FCTC)** Cluster centers are fixed and at most α number of cluster changes are allowed in the entire time series data.	**Constrained Time Series Clustering (CTSC)** Cluster centers can change and at most α number of cluster changes are allowed in the entire time series data.

Table 1.1: A Summary of Purposed Different temporal clustering schemes that arise from the dynamic cluster center choice (column) and dynamic assignment choice (rows).

Indeed, these six algorithms effectively bridge the gaps existing between TSC and TLA. By employing constraint optimization and drawing upon the optimal K-Centers formulation designed for non-temporal clustering [13], Tosanwumi formalized six clustering schemes within the design space, as illustrated in Table 1.1, through a Mixed Integer Linear Program (MILP) formulation. Significantly, the solution derived from this approach is consistent, reproducible, and insensitive to initialization, in stark contrast to previous temporal clustering algorithms. Moreover, it achieves global optimality, marking a substantial advancement in addressing the shortcomings of earlier methods.

1.3 Project Objective

The six clustering schemes outlined in the previous section form a robust toolkit for temporal data analysis, holding potential applications across various disciplines. However, the introduction of this newly proposed algorithm prompts further exploration to address specific research problems. In collaboration with Tosanwumi, this project will focus on two main aspects:

Using synthetic data to conduct a performance comparison:

- **RQ1**: Do optimal schemes with fixed cluster assignment outperform popular existing approaches, such as Time Series Clustering (TSC) and Dynamic Time Wrapping [14] (DTW), in terms of purity and the maximum distance between a cluster and a time series entity with known clusters?
- **RQ2**: Compared to existing Temporal Label Analysis (TLA) approaches, can schemes with label changes (i.e., FUTC, UTC, FCTC, CSTC) correctly identify known cluster label changes?
- **RQ3**: Among all variations, such as the sum of distances vs. k-center objective, hard constraints vs. Lagrangian dual, and different techniques, which algorithm proves to be the most efficient?

Using real-world datasets to discover potential problems and solutions:

- **RQ4**: What are potential issues in existing approaches that might lead to inaccurate results or increase difficulties in analysis?
- **RQ5**: Within these identified issues, does the proposed optimal temporal clustering framework demonstrate improved results or approaches?
- **RQ6**: Is the guarantee optimality in our proposed framework showing real clustering differences and new insight in real-world examples, instead of different values of cluster centers?

The key insight of this project lies not only in providing a direct performance comparison between previous temporal clustering algorithms and the newly proposed schemes but also in conducting a thorough analysis of several real-world datasets using these optimal temporal clustering algorithms. These datasets encompass daily climate data spanning from 2018 to 2022 [15], covering 192 locations worldwide, including major cities like Ottawa and Berlin, as well as smaller islands and rural research stations. Additionally, Toronto census data [16] from 1996 and 2021, based on Forward Sorted Area (FSA) and neighbourhoods, offer insights into hidden trends in the percentage of population in occupation in art culture, as well as the stability and volatility of different areas in Toronto. This multifaceted approach allows for a comprehensive understanding of the strengths and unique contributions of the proposed temporal clustering framework in real-world scenarios.

Chapter 2

Related Work

Prior research in temporal clustering has focused on changing cluster definition with fixed assignment (standard time series clustering). Most research on this temporal clustering scheme is mainly centred on trading off between time series representation (mainly through dimensional reduction) and similarity metric [17].

Dimensionality reduction leads to problems in the choices of time series data representation, some popular approaches including Piecewise Aggregate Approximation (PAA) [18] involve reducing the dimension of time series data by taking the mean of values in a moving non-overlapping window with a fixed size [18]. And its adaptive variant called Adaptive Piecewise Constant Approximation (APCA) [19, 20] involves calculating the mean [19] using deep autoencoder networks with varying window size [17]. The limitations of these methods are that they can lead to the loss of important features such as the decay of long-term temporal correlations [17]. A possible solution for this limitation is used deep auto-encoder networks to reduce the dimension of time series data before clustering [17]. However, deep autoencoder networks are computationally expensive and may give sub-optimal results.

The common similarity metrics to cluster time series include Euclidean distance [21, 22] and Dynamic Time Warping (DTW) [23]. While DTW can be used for both whole-time series clustering and shape clustering, it is intractable to fit it into a MILP framework (which has the advantage of guaranteed optimality). On the other hand, Euclidean distance has less meaningfulness than L_1 norm in high dimensional space [24] such as in time series clustering.

In this project, we conducted case studies in two study fields, census data and climate data. Census data clustering, the identification of potential groups in census data without prior knowledge about the population, serves as a vital tool for sociological and geological researchers [25]. However, most current approaches predominantly focus on either spatial clustering of sample distribution [26], encompassing only a single timestamp [27], or aggregating multiple features into time series within a single dimension [28]. The clustering of census data using temporal datasets with multiple timestamps and features, known as Census Temporal Data Clustering, remains largely unexplored.

Therefore, We see that there is a gap in having efficient, scalable, algorithms for temporal clustering to create a guaranteed optimal solution for many potential real-world applications with temporal clustering data.

Chapter 3

Methology

The key approach for performance comparison and analysis of real-world datasets involves employing MILP formulations for the six optimal temporal clustering schemes outlined in Table 1.1 by Jolomi Tosanwumi. This section will begin by introducing the MILP objective of these schemes, followed by a detailed discussion of the formulations for each proposed scheme in Table 1.1.

3.1 The Objective of the MILP Formulation for Proposed Schemes

Consider a time series entity X_{li} in the time series data $X \in \mathbb{R}^{l \times n \times d}$, where l is the number of time steps, n is the number of time series entities, and d is the number of dimensions of each time series entity. If we aim to assign each time series entity to one of k clusters, then we define the following:

- X_{ti} ($\in \mathbb{R}^d$) is a vector of the values of observed time series i at time t
- C_{ij} is an indicator variable that shows if time series i is assigned to cluster j, where $i \in N$, $j \in J$, $N = \{1...n\}$, and $J = \{1...k\}$.
- Z_j ($\in \mathbb{R}^d$) is the cluster center of cluster j
- $\mathcal{E}_{ti}$ is a decision variable whose optimal value will be the distance between time series i and its assigned cluster at time t, where $T = \{1...l\}$.

Using this setting, Tosanwumi formulated temporal clustering as a MILP problem for the different clustering schemes mentioned in Table 1.1 by employing a temporal extension of k-medians clustering [29]. The primary objective of the k-median clustering problem is to minimize the sum of L_1 distance between all data points and their cluster(s), as defined by:

$$\min_{C,Z,\mathcal{E}} \sum_{t \in T} \sum_{i \in N} \mathcal{E}_{ti} \tag{3.1}$$

An optimization problem using 3.1 as an objective often yields a manageable number of variables to optimize in general cases. However, this assumption is not typically true in problems involving temporal data. When there is a large number of time steps l or a considerable amount of time series n, in both cases, the size of data points to be considered $n \times l$ will dramatically increase, resulting in an increase in computational time and potential scalability issues.

To mitigate the challenges associated with scalability, Tosanwumi modified the objective from 'sum of distances' to 'maximum distance' in the MILP formulations. Instead of minimizing the sum of L_1 distances between all data points and their cluster(s), the new objective only minimizes the *maximum* cluster violation (distance) in a variation of k-centers [13].

$$\min_{C,Z} \max_{t\in T} \max_{j\in J} \max_{i\in N} \|X_{ti} - Z_{tj}\|_1 C_{tji} \tag{3.2}$$

The $L1$ distance is multiplied by C_{tij} to ensure that we are only considering the cluster a time series is assigned to at each point in time. This new objective, also known as k-center, effectively reduces the number of data points from $n \times l$ to 1 without compromising the goal of global optimality. As one of the objectives of this project, we will examine the performance difference between k-median and k-center objectives in the future section. To further transform 3.2 from a min-max MILP into a pure MILP, Tosanwumi introduced a new variable $\mathcal{E}$, such that $\mathcal{E}$ will equal the maximum cluster distance at optimality.

$$\min_{C,Z,\mathcal{E}} \mathcal{E} \tag{3.3}$$

Employing this objective, we will present a concise description and the MILP formulation for each of the proposed schemes in Table 1.1.

3.2 Fixed Time Series Clustering (FTSC)

In FTSC, both cluster centers and label assignments of samples remain fixed across all timestamps $t \in T$. Figure 3.1 is a visualized example of FTSC.

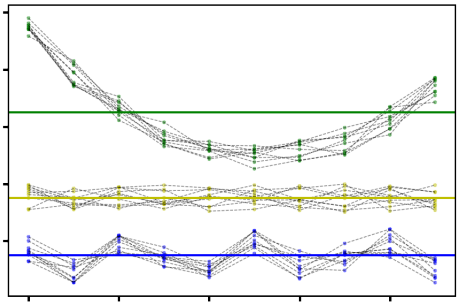


Figure 3.1: Clustering Schemes Visualization of FTSC. The dashed lines represent time series samples. while the solid lines are three cluster centers.

As the cluster centers are fixed throughout all t, they are depicted as straight lines in Figure 3.1. The cluster label assignments (the color of scatter) remain consistent across all t for each sample. Aside from 3.3, to achieve this goal using a pure MILP formulation, we also need a set of constraints to regulate the changes.

$$C_{ij} \in \{0,1\} \; \forall \; i \in N \; j \in J; \sum_{j\in J} C_{ij} = 1 \; \forall \; i \in N \tag{3.4}$$

Equation 3.4 introduces $n \times k$ binary indicator variables serving as the cluster assignment for each sample. c_{ij} equals 1 if sample i is in cluster j, and 0 otherwise. As the cluster assignment is fixed throughout all t in FTSC, for each sample, only one indicator variable is needed across all t. In addition, $\sum_{j\in J} C_{ij} = 1$ ensures

that each sample is assigned to only one cluster.

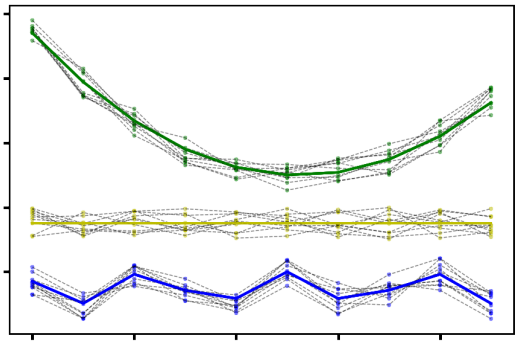
$$C_{ij} = \begin{cases} 1, & \text{if } \|X_{ti} - Z_j\|_1 \leq \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \quad \forall\, t \in T, i \in N, j \in J; \mathcal{E} \in \mathbb{R}; Z_j \in \mathbb{R}^d \,\forall\, j \in J \tag{3.5}$$

To convert from the minimax MILP 3.2 to the pure MILP 3.3, 3.5 includes $l \times n \times k$ logical constraints ensuring that the maximum violation $\mathcal{E}$ is greater than the L_1 distances between all data points and their cluster centers. Since the cluster centroid is also fixed through t in FTSC, only the simplified version of z_{tj}, z_j, is needed for the L_1 distance. The completed MILP formulation of FTSC is

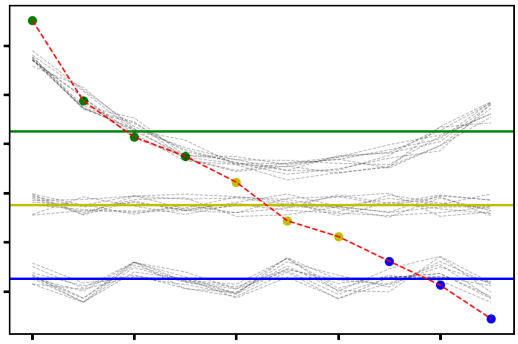
$$\begin{aligned} \underset{C,Z,\mathcal{E}}{min} \quad & \mathcal{E} \\ \text{s.t.} \quad & C_{tij} = \begin{cases} 1, & \text{if } \|X_{ti} - Z_j\|_1 \leq \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \\ & \forall\, t \in T, i \in N, j \in J \\ & \textstyle\sum_{j \in J} C_{tij} = 1 \,\forall\, t \in T, i \in N; \\ & C_{tij} \in \{0,1\} \,\forall\, t \in T, i \in N j \in J; \\ & Z_{tj} \in \mathbb{R}^d \,\forall\, t \in T, j \in J; \mathcal{E} \in \mathbb{R} \end{aligned} \tag{3.6}$$

3.3 Standard Time Series Clustering (STSC) and Fixed Unconstrained Temporal Clustering (FUTC)

The difference between **STSC** and **FTSC** is **STSC** permits each cluster center to dynamically change across timestamps t, $\forall t \in T$. This leads to the curved cluster centers that align with the trend of samples, shown as solid lines in Figure 3.2a. Conversely, the cluster assignment remains fixed for each sample. It is illustrated in Figure 3.2a as consistent colour scatter across t within each sample.



(a) STSC Clustering Schemes Visualization



(b) FUTC Clustering Schemes Visualization

Since the value of the cluster j's center at timestamp t, Z_{tj}, is not fixed across t, we need to adjust the portion of the MILP formulation responsible for calculating the L_1 distance between data points and their

centers to accommodate this change.

$$C_{ij} = \begin{cases} 1, & \text{if } \|X_{ti} - Z_{tj}\|_1 \leq \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \quad \forall\, t \in T, i \in N, j \in J; \mathcal{E} \in \mathbb{R}; Z_{tj} \in \mathbb{R}^d \; \forall t \in T, j \in J \tag{3.7}$$

This change enables the calculation of the L_1 distance between sample i and the current value of its center at every timestamp t. While other parts of **STSC**'s MILP formulations will be the same as the previous **FTSC**'s MILP formulation 3.6.

In contrast, **FUTC** allows dynamic changes in cluster assignments for all samples but not cluster centers. If, at any timestamp, a sample i exhibits a L_1 distance smaller than the current cluster's center with other cluster centers, **FUTC** will assign this sample to the new cluster, resulting in a change in cluster assignment. This behavior is exemplified in Figure 3.2b, where one of the samples in the red dash line transitions from the green cluster to yellow and then blue. This change occurs because its value possesses a smaller L_1 distance to the yellow (blue) cluster center than to its previous green cluster center. However, the cluster center remains fixed across all timestamps in **FUTC** (straight solid line). Similarly, we need to modify the MILP formulation to accommodate this change.

$$C_{tij} \in \{0, 1\} \; \forall\, t \in T \; i \in N \; j \in J; \sum_{j \in J} C_{tij} = 1 \, \forall\, t \in T \; i \in N \tag{3.8}$$

It's worth noting that, since cluster assignment is dynamic, using only one binary indicator variable for each sample at each cluster is not sufficient. Therefore, 3.8 introduces t indicator variables for each timestamp in each sample, resulting in a total of $l \times n \times k$ binary indicators. When the size of T becomes relatively large, the number of constraints will also increase dramatically. Currently, there is limitation in the scalability of **FUTC**.

3.4 Unconstrained Temporal Clustering (UTC)

UTC can be understood as a combination of **FUTC** and **STSC**, which indicates both cluster centers and cluster assignments can change across all timestamps t. without any limitation. This results in both curved cluster center (solid lines) and sample i (red dash line) which change cluster assignments, as shown in the visualization of UTC Figure 3.3.

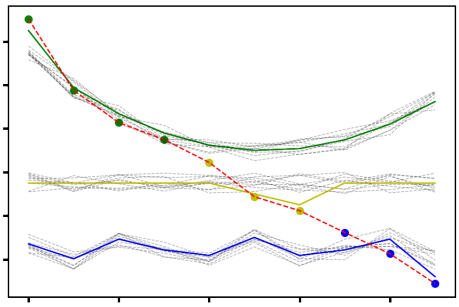


Figure 3.3: UTC Clustering Schemes Visualization

The MILP formulation of **UTC** adapts the modification made for both **STSC** 3.7 and **FUTC** 3.8 since both cluster centers and assignments are dynamic. The complete formulation will be

$$
\begin{aligned}
\min_{C,Z,\mathcal{E}} \quad & \mathcal{E} \\
\text{s.t.} \quad & C_{tij} = \begin{cases} 1, & \text{if } \|X_{ti} - Z_{tj}\|_1 \leq \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \\
& \forall\, t \in T, i \in N, j \in J \\
& \textstyle\sum_{j \in J} C_{tij} = 1 \,\forall\, t \in T, i \in N; \\
& C_{tij} \in \{0,1\} \;\forall\, t \in T, i \in N j \in J; \\
& Z_{tj} \in \mathbb{R}^d \;\forall\, t \in T, j \in J; \mathcal{E} \in \mathbb{R}
\end{aligned} \tag{3.9}
$$

These four clustering schemes cover the design space regarding whether Z or C can dynamically change over T, effectively filling the gaps between previous approaches. Tosanwumi also extended these schemes to solve other useful problems.

3.5 Bounded Cluster Assignment Change

Understanding which time series samples are more prone to changing clusters can be valuable in the study of clustering stability and volatility [30]. To address this need, the previous cluster schemes can be modified to limit the maximum number of changes in cluster assignments for each sample.

Building on the MILP formulations in the previous sections, there are two approaches to this problem. The first approach involves adding $l \times n$ indicator variables y_{ti}, defined in 3.10, to identify whether time series i is assigned to the same cluster in the tth and $(t+1)$th time steps.

$$
\begin{aligned}
y_{ti} = \begin{cases} 1, & \text{if } \sum\limits_{j \in J} j \cdot C_{t,i,j} = \sum\limits_{j \in J} j \cdot C_{t+1,i,j} \\ 0, & \text{otherwise} \end{cases} \\
\forall\, t \in T \backslash \{l\}, i \in N, j \in J
\end{aligned} \tag{3.10}
$$

Using y_{ti}, we introduced a hard constraint to limit the number of cluster label changes is not less than α.

$$
\sum_{t \in T \backslash \{l\}} \sum_{i \in N} y_{ti} \geq (l-1) \cdot n - \alpha \tag{3.11}
$$

This approach is known as the 'hard constraint' approach. It directly regulates the number of allowed cluster changes. This constraint can be added to any of the four clustering schemes' MILP formulations in the previous sections to explicitly adjust α, the parameter that regulates the maximum number of changes, to a desired amount.

However, in most cases, a suitable α is unknown. Alternatively, we can also consider a soft-constrained dynamic cluster assignment problem by relaxing the constraint and penalizing violation of the constraint in a Lagrangian dual formulation. This will result in modifying the objective functions 3.3to form the respective Lagrangian dual formulations. The objective function of the Lagrangian dual problem is shown below:

$$
\min_{C,Z,\mathcal{E}} \quad \mathcal{E} + \lambda \cdot \frac{||\max(X) - \min(X)||_1}{n(l-1)} \sum_{t \in T \backslash \{l\}} \sum_{i \in N} (1 - y_{ti}) \tag{3.12}
$$

where

- λ is the Lagrangian multiplier that penalizes cluster label changes.

- $\max(X), \min(X) \in X^d$ are vectors of the maximum and minimum values in each dimension of the data X respectively. (The L_1 distance between them is a measure of the maximum possible distance between an observation and its cluster center).
- $n(l-1)$ is the maximum possible number of cluster changes.
- $\frac{\max(X)-\min(X)}{n(l-1)}$ is a scaling term to scale the number of cluster changes to that of $\mathcal{E}$

As λ is decreased from ∞ to 0, the number of cluster changes will increase. This is important for finding time series entities that tend to change clusters first. However, unlike the hard-constrained version which explicitly specifies the number of allowed changes α, one must perform a binary search on λ to find a setting that corresponds to exactly α changes.

Applying either of these two approaches to **FTSC** results in Fixed Constrained Temporal Clustering (FCTC). FCTC allows for bounded changes in cluster assignments while maintaining fixed cluster centers across t. On the other hand, applying these approaches to **STSC** yields Constrained Time Series Clustering (CTSC). CTSC permits both bounded changes in cluster assignments and dynamic changes in cluster centers.

Chapter 4

Performance Evaluation

This chapter undertakes the evaluation of the proposed schemes' performance, utilizing a synthetic dataset to address the initial three research questions outlined in the project objectives. These questions are further dissected into comparisons between the proposed schemes and existing approaches.

- **Comparison 1**: Contrasting the proposed optimal schemes, allowing dynamic center but fixed assignment, with existing Time Series Clustering (TSC) and Dynamic Time Wrapping (DTW) approaches. The evaluation criteria include purity and the maximum distance among clusters' centers and their time series entities.
- **Comparison 2**: Assessing the proposed optimal schemes, permitting dynamic label changes, in comparison with Time Label Analysis (TLA) concerning their effectiveness in detecting existing label changes within samples.

Additionally, comparisons within the proposed schemes are conducted based on runtime:

- **Comparison 3**: Two proposed MILP objectives: k-median 3.1 versus k-centers 3.3.
- **Comparison 4**: Two proposed bounded label change approaches: Hard Constraint 3.11 versus Lagrangian Dual 3.12.
- **Comparison 5**: Different proposed techniques in speed optimization.

The primary goal of this performance evaluation is to establish a direct and comprehensive comparison between Tosanwumi's proposed schemes and existing approaches. Importantly, it aims to use the evaluation result to unveil the unique potential applications of the proposed schemes in real-world scenarios.

4.1 Synthetic Data

To mitigate the potential bias stemming from real-world datasets, which may impact the fairness and robustness of the evaluation process [31], we constructed a synthetic dataset illustrated in Figure 4.1 for performance comparison. This dataset encompasses three distinct groups of time series data, each comprising 10 timestamps in every sample, showcasing unique trends and patterns. The initial three samples are generated using the subsequent mathematical functions:

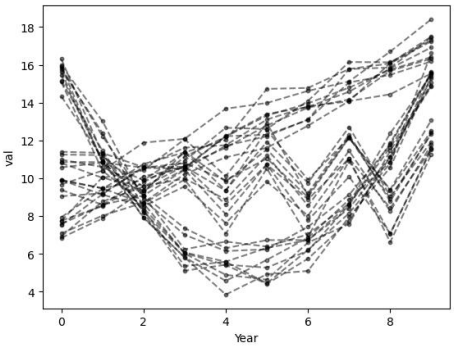


Figure 4.1: 30 Synthetic Time Series samples from three functions with Gaussian noises

$$\text{Sample 1: } t_i = 0.5x^2 - 4.5x_i + 15 \; ; x \in \{1, ..., 10\} \tag{4.1}$$

$$\text{Sample 2: } t_i = x_i + 8 \; ; x \in \{1, ..., 10\} \tag{4.2}$$

$$\text{Sample 3: } t_i = 2\sin(3x_i) + 10 \; ; x \in \{1, ..., 10\} \tag{4.3}$$

Subsequent samples are reproduced by introducing distinct Gaussian noises, affording us control over the size and randomness of the samples. This dataset serves as a comprehensive platform for comparing purity, maximum distance in **Comparison 1**, and efficiency in **Comparison 3, 4, and 5** across various methodologies.

To assess the performance in identifying changes in cluster labels within the scope of **Comparison 2**, we created an additional synthetic dataset (Figure 4.2). This dataset features randomly generated time series samples exhibiting relatively stable trends, with no discernible differences between timestamps. Two samples were manually introduced with substantial changes across timestamps.

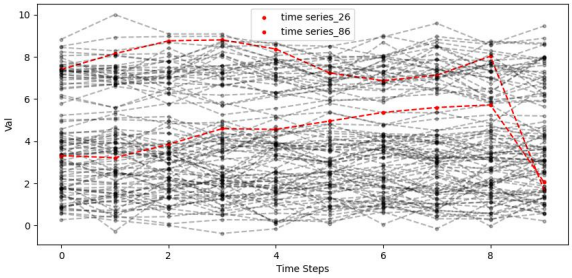


Figure 4.2: Time series in red color represent two manually added samples with changes across timestamps.

By evaluating the capability of different approaches to distinguish these manually added samples, we can gauge the effectiveness of the proposed schemes in identifying cluster label changes.

4.2 Performance in Fix Cluster Assignment Schemes

The proposed Standard Time Series Clustering (STSC) 3.7 is utilized in **Comparison 1 and 3**. STSC is chosen over other schemes because it specifically allows changes in the cluster center but not assignment changes, aligning with the other two existing approaches in **Comparison 1** to enforce fairness in the comparison.

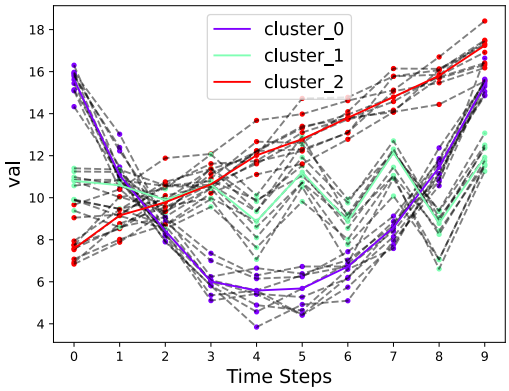
To conduct **Comparison 1**, synthetic data from Figure 4.1 is used with different noise levels achieved by multiplying the Gaussian noise at different scales during the data generation processes. Table 1.1 presents the purity and maximum distance between cluster centers for existing Time Series Clustering approaches, namely K-means clustering (km-TSC), K-means Dynamic Time Warping (km-DTW), and STSC. Since the synthetic data comprises univariate time series data, the L_1 distance of samples equals the L_2 distance. Therefore, the maximum distance between cluster centers and entities is measured using only L_1 distance.

Noise Factor	Purity Score			Max L_1 Distance		
	km TSC	km DTW	STSC	km TSC	km DTW	STSC
0.500	**1.0**	**1.0**	**1.0**	1.858	2.823	**1.586**
0.875	**1.0**	**1.0**	**1.0**	2.560	4.300	**2.170**
1.250	**1.0**	**1.0**	**1.0**	3.354	5.289	**2.793**
1.625	**1.0**	**1.0**	**1.0**	4.413	6.022	**3.416**
2.00	**1.0**	**1.0**	0.967	5.485	7.880	**4.131**

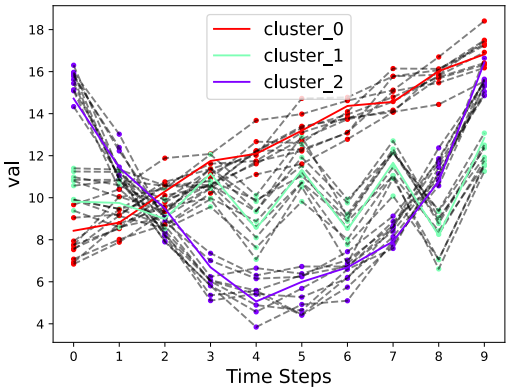
Table 4.1: Synthetic Evaluation Results of Different Approaches for Temporal Clustering

The results indicate that the proposed optimal STSC scheme outperforms both existing approaches in reducing the maximum distance between cluster centers and entities while maintaining similar purity. The margin in maximum distance between STSC and existing approaches increases as the noise level in the dataset increases, indicating that STSC is more resistant to noise in the dataset.

To address **Comparison 3**, we first confirmed two objectives of the optimization problem: the sum of distance (k-median) in Figure 4.3a and maximum distance (k-center) in Figure 4.3b can successfully cluster all the synthetic samples into their original groups using STSC scheme.



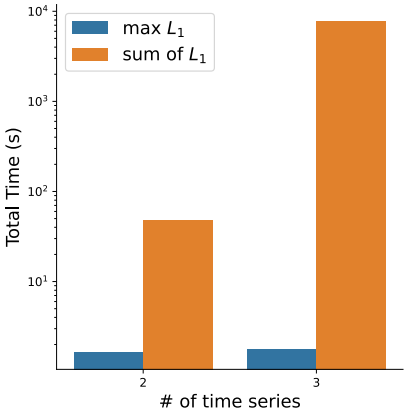
(a) STSC Clustering Result of k-median



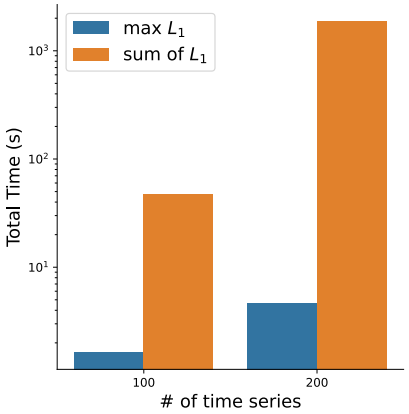
(b) STSC Clustering Result of k-center

Then, we compared the runtime (in seconds) difference between the two objectives in different sample sizes by generating more samples using Gaussian noise into three groups, and different output clusters. The maximum distance objective outperforms the sum in two different numbers of clusters (Figure 4.4a) and different sample sizes (Figure 4.4b).

The result suggested that converting from k-median to k-center to reduce the total constraints size can significantly improve the model runtime without sacrificing the optimality of the result. Validating the effectiveness of Tosanwumi's approach. We can use the k-center objective in real-world applications to reduce the



(a) Runtime of two objectives in 2 and 3 clusters settings



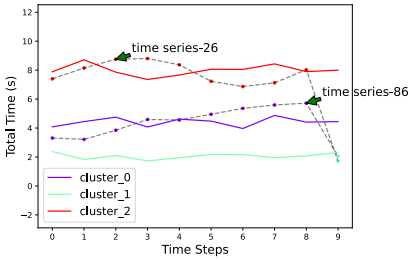
(b) Runtime of two objectives in 100 and 200 sample sizes

computational cost of the model.

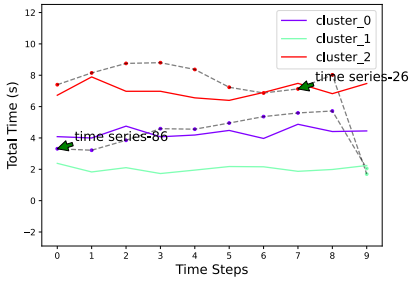
4.3 Performance in Bounded Label Change Schemes

The Constrained Time Series Clustering (CTSC) scheme is employed for both changes in centers and bounded changes in assignment. This makes it a suitable scheme for comparing the ability to determine label changes with dynamic cluster center, a capability not present in TLA, as highlighted in **Comparison 2**. Furthermore, CTSC, with one of the highest numbers of total constraints, enables an exploration of runtime differences between bounded label change approaches in **Comparison 4**.

In **Comparison 2**, CTSC provides extensive control over time series mining of dynamic entities in both hard constraint approaches (Figure 4.5a) and Lagrangian relaxation approaches (Figure 4.5b). This flexibility is notably absent in 'Temporal Label Analysis (TLA),' which re-clusters without a time domain and is constrained to fixed cluster centers.



(a) **Hard Constraint** Two Entities identified.



(b) **Lagrangian Relaxation** Two Entities identified

The insights gained from detecting label changes with a dynamic cluster center can expand our design space in real-world applications. In the following section, we will discuss the side-by-side differences found in the Toronto census trend using both fixed-center and dynamic-center approaches, providing unique insights found only in CTSC.

In **Comparison 4**, we compared between the Hard-Constraint 3.11 and Lagrangian Dual 3.12 approaches with different maximum allowances in cluster assignment. The results (Figure 4.6) show a dramatic increase

in runtime when using the Lagrangian Dual approach.

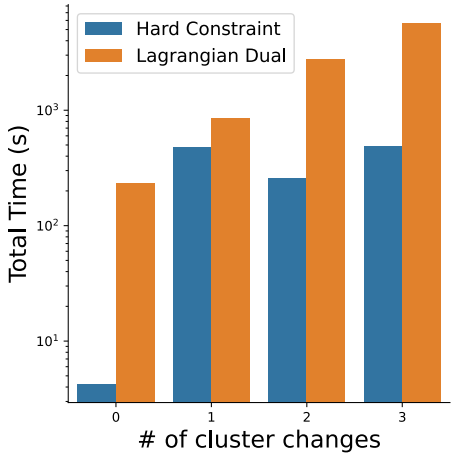
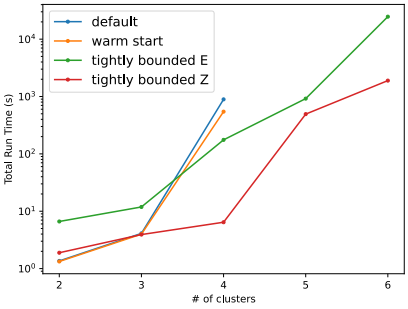


Figure 4.6: The runtime (log scale) difference between hard constraints and Lagrangian Dual using 3 clusters **CTSC** with different maximum allowances in cluster assignment.

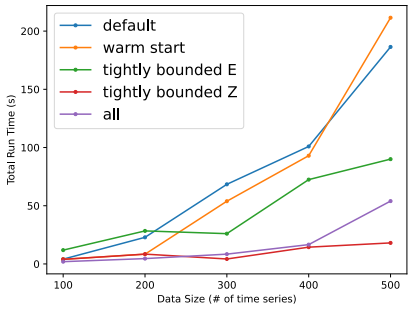
This unexpected insight contradicts established results [32] demonstrating the efficacy of Lagrangian relaxations in improving the performance of large-scale mathematical programming questions. It approximates a challenging optimization problem by employing multiple simpler problems. However, this does not hold in the proposed optimal temporal clustering schemes. Therefore, in real-world applications, attention should be given not only to the Lagrangian dual approach but also to the hard constraint approach, especially when computational time is a critical consideration.

4.4 Performance in Different Techniques

We employed the same metrics and synthetic data to compare runtime differences between techniques detailed in the previous section, including warm start, 'tightly bounded $\mathcal{E}$,' and 'tightly bounded Z.' In various



(a) Runtime among techniques in different cluster settings



(b) Runtime among techniques in different sample sizes

sample sizes (Figure 4.7b), we observed that schemes using the 'tightly bounded Z' technique exhibited a significant reduction in runtime compared to other techniques as the sample size increased. Conversely, schemes without any additional techniques (base) and those using 'warm-start' had the longest runtime.

A similar scenario unfolded in the experiment with different cluster sizes (Figure 4.7a). 'Tightly bounded

Z' continued to dominate performance among all techniques. It is noteworthy that the base and 'warm-start' schemes were unable to finish within a feasible runtime after the 4-cluster setting.

Evaluating the performance of different techniques in speed optimization provides us with guidelines on which techniques should be employed when designing a suitable scheme for real-world applications. The significant runtime reductions observed in the 'tightly bound Z and E' techniques can assist in expanding the scope of our analysis to higher cluster settings, larger sample sizes, and additional features simultaneously.

Chapter 5

Census Data

5.1 Introduction

A census is a systematic and comprehensive collection of data related to a specific population or area, typically conducted at a fixed frequency across years and published as a series of tables [33]. By integrating historical census data with various features into a consolidated dataset, the temporal census data created a profile across time specific to a particular population. Analyzing and interpreting census data over time can unveil valuable insights and trends within the studied population. Additionally, connecting temporal census data with Geographical Information Systems (GIS) allows the creation of geo-referenced data [27], facilitating the examination of migration statistics, demographic changes [34], and workforce allocations.

Census Data Clustering, the identification of potential groups in census data without prior knowledge about the population, serves as a vital tool for sociological and geological researchers. This case study will utilize Toronto Census data from 1996 to 2001, encompassing 7 features and 95 samples. Due to the advantages of the proposed framework, which spans the entire design space between dynamic and fixed cluster centers and assignments, we can uncover trends in multiple features originally concealed in the results of existing approaches. Furthermore, it facilitates corrections for misinterpreted insights in current approaches.

The result of this case study will answer **RQ4** and **RQ5** in the project objective, by initially pinpointing potential issues if using current TSC and TLA approaches for Census Temporal Data Clustering, and proposing potential improvements to address these challenges through the implementation of the optimal framework. The study will delve into both the concealed trends within each cluster by scrutinizing the aggregation of its samples and unravelling patterns in specific samples through a stability analysis.

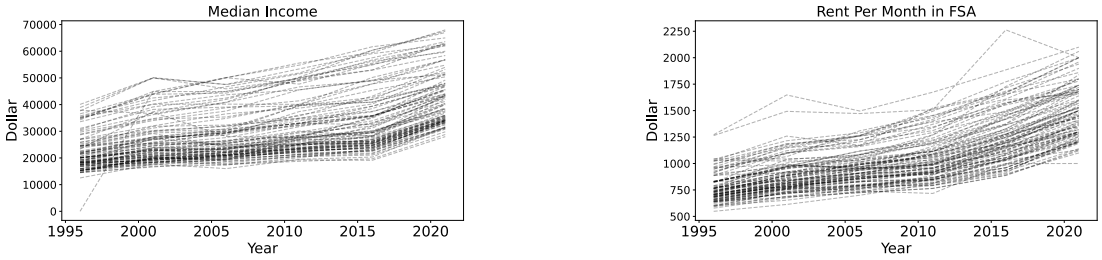
5.2 Datasets used

The temporal data utilized in this case study comprises Toronto Census Data spanning the years 1996 to 2021, encompassing 95 Forward Sortation Areas (FSAs). This dataset was obtained from the Canadian Census Analyser [35] in the CHASS Datacenter. Given the frequency of the Canadian census population, conducted once every 5 years, each sample in this temporal dataset comprises 6 timestamps.

The original census data contains an extensive array of features. However, not all features could be included. Similar to climate data considerations, an abundance of features would significantly amplify the number of constraints in MILP, leading to infeasible runtimes using current methodologies. Additionally,

including too many features would impede the ability to discern which subset influences the clustering center, thereby complicating the analysis. Consequently, for this case study, seven features were selected to represent the Toronto Census trends. These features include 'Percentage of FSA Population in Total Toronto Population,' 'Percentage of Visual Minorities in FSA Population,' 'Percentage of the Population Commuting by Walk,' 'Percentage of the Population with an Occupation in Art and Culture,' 'Median Income,' and 'Average Rent per Month.' The resulting temporal dataset consists of 95 samples, each containing 6 timestamps across the seven selected features.

Within this set of features, 'Median Income' and 'Average Rent per Month' (see Figure 5.1) exhibit a general increasing trend over time across all samples. This upward trajectory is attributed to various environmental factors, including inflation [36]. Importantly, the presence of this overarching increasing trend makes it challenging to uncover any hidden trends within the samples across the years.



(a) Median Income

(b) Average Rent per Month

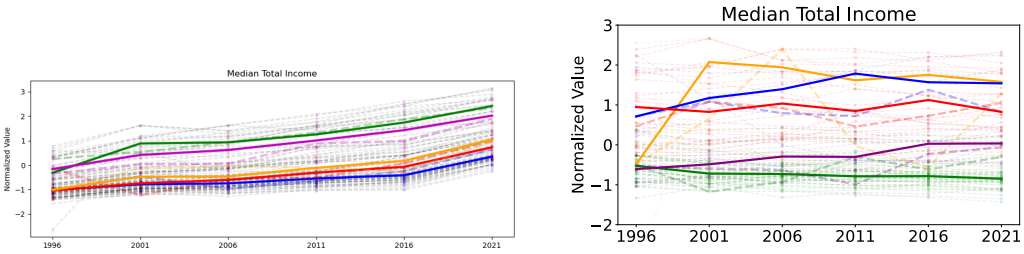
Figure 5.1: Increased trends among all-time series in feature median income and average monthly rent

To address this issue, we employed z-score normalization, individually normalizing the values of each feature for every year. This normalization ensures that all timestamps and features share a standardized scale. Experimental results validate the efficacy of this solution in enhancing the visibility of latent trends within the temporal data. For instance, when utilizing STSC with 5-cluster settings, the results (Figure 5.2a) obtained from non-normalized data only reveal a slightly steeper slope within the 'green' encoded cluster between 1996 and 2001. In this scenario, other trends are concealed within the overarching increasing trend. In contrast, the results after normalization (Figure 5.2b) exhibit distinguishable trends in both 'purple' and 'yellow' encoded clusters. Notably, the normalized data reveals a drop in median income after 2001 within the purple encoded clustering, a trend completely obscured in the earlier results. This underscores the effectiveness of z-score normalization in uncovering hidden patterns within the temporal data.

5.3 Process

As articulated earlier, the primary objective of this case study is to discern any distinctive trends revealed by our proposed schemes that were not evident in previous approaches. However, a majority of existing temporal clustering approaches employ k-mean clustering, which employs a different cluster and error definition than the k-center clustering in our proposed optimal schemes.

To ensure a fair comparison without compromising the demonstrated advantages of k-center in runtime reduction, we opted to utilize the Standard Time Series Clustering (STSC) to illustrate the results of the existing Time Series Clustering (TSC) approach. This choice is justified by the fact that both STSC and TSC allow dynamic cluster centring while fixing cluster assignments. Similarly, we employ the Fixed Unconstrainted



(a) Before Normalization

(b) After Normalization

Figure 5.2: Median Income from STSC in 5-cluster setting using data from both before and after normalization. The solid line shows the median of samples in each cluster, and the dashed line shows the corresponding cluster centers.

Time Clustering (FUTC) to substitute the results of the existing Time Label Analysis (TLA) approach.

Since k-center clustering relies only on the maximum difference between samples and their cluster center, this results in clusters' centers being easily affected by outliers. To illustrate the trend shown in samples more effectively, we used the median to aggregate the value of samples within each cluster. For example, in Figure 5.2b, the cluster center of the pink encoded cluster (pink dash line) is extremely volatile, with a global maxima point in 2006. However, the median of samples (pink solid line) shows a relatively stable decreasing trend after 2001, which is completely contradicted by the trend shown in the cluster center. In addition, we choose median, instead of mean aggregation to be more resistant to potential skewness shown in samples (Figure 5.3)

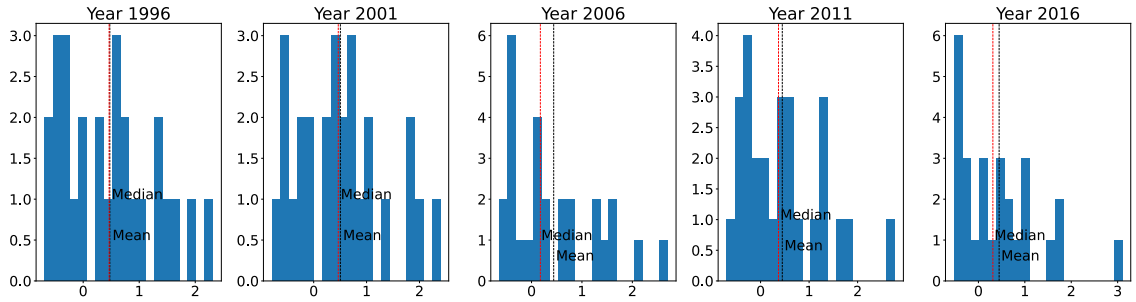


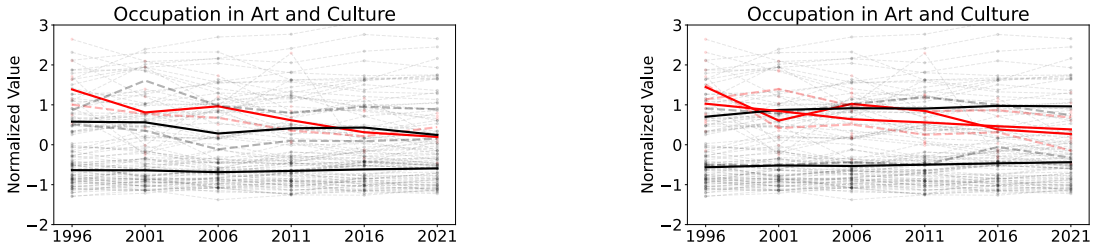
Figure 5.3: An illustration of skewness in the sample, red dash line is the median of the sample, black dash line is the mean. In skewed data, such as in 2006, the median will be a better representation than the mean.

5.4 Trends of Occupation in Art and Culture

5.4.1 Problem in existing approaches

The percentage of the total workforce engaged in the occupation of art and culture serves as a significant metric, reflecting the existing correlation between education and the economic landscape of a given area [37]. Examining the clustering results of the percentage of art and culture occupation in each workforce using STSC with a 3-cluster setting (Figure 5.4a), we observe a decreasing trend (depicted in red) in the median of samples within one of the clusters. Conversely, the other clusters (depicted in black) exhibit a relatively flat curve. Similar findings emerge in the 4-cluster setting (Figure 5.4b), where two clusters demonstrate

decreasing trends (in red) while the remaining clusters exhibit stable, flat trends (in black). Both sets of results converge on a consistent conclusion: the overall percentage of the workforce involved in the occupation of art and culture is decreasing over the years in Toronto.



(a) 3-cluster setting

(b) 4-cluster setting

Figure 5.4: the STSC results for the percentage of the workforce engaged in the occupation of art and culture under both the 3-cluster and 4-cluster settings. The solid line denotes the median of samples within the respective cluster. Conversely, the dashed line represents the cluster center.

However, the analysis of actual census data reveals a different outcome. In Figure 5.5, the total percentage of the workforce in the occupation of art and culture exhibits a generally increasing trend over the years, with only a slight drop in 2011. The conclusion of a decreasing trend, drawn solely from the clustering results of the existing TSC approach, contradicts the ground truth data.

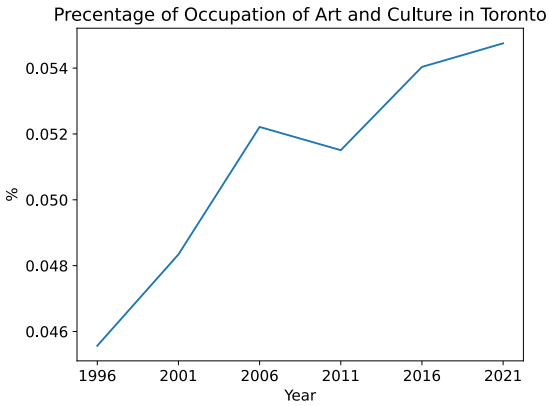
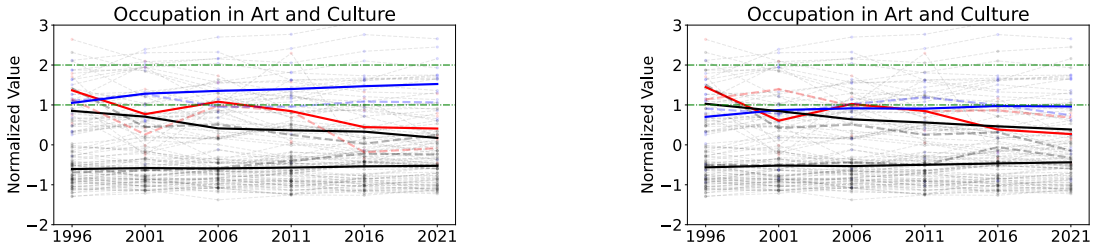


Figure 5.5: A increased trend shown in percentage of art occupation in Toronto

5.4.2 Clustering Result using CTSC

The proposed CTSC scheme expands upon the existing design space by allowing both changes in cluster assignments and dynamic cluster centers. Examining the clustering results of the same feature, the percentage of the workforce engaged in the occupation of art and culture, but using CTSC with a 4-cluster setting (Figure 5.6a), the blue-encoded cluster eventually exhibits the increasing trend observed in the ground truth data. Meanwhile, the cluster with a decreasing trend (in red) is still present in the results. As a reference, there is no discernible trend in the blue-encoded cluster (containing a similar subset of samples) in STSC (Figure 5.6b).

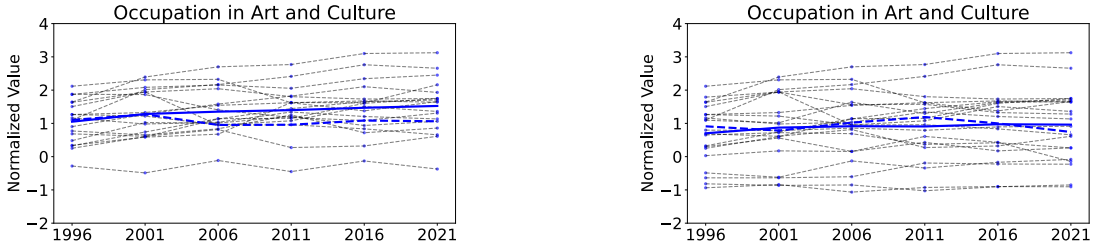


(a) CTSC in 4-cluster setting

(b) STSC in 4-cluster setting

Figure 5.6: The results for the percentage of the workforce engaged in the occupation of art and culture from both CTSC and STSC, utilizing a 4-cluster setting, Notably, the blue-encoded cluster in CTSC shows an increasing trend across the years, whereas it appears relatively flat in STSC. For visual comparison, the green dot-dash lines serve as a reference range between 1 and 2.

In Figure 5.7, both schemes exhibit similar cluster centers (blue dash line) for the blue-encoded cluster. However, notable differences emerge in the distribution of samples (black dash line) clustered in the blue-encoded clusters between the CTSC and STSC schemes. CTSC, depicted in Figure 5.7a, displays a more concentrated distribution around the cluster center, with the majority exhibiting an increasing trend over the years. Conversely, samples from STSC, as illustrated in Figure 5.7b, contain more samples that remain relatively stable over time, lacking a significant increasing trend compared to other samples. While STSC can cluster samples with increasing trends together, the presence of samples with different trends complicates correct interpretations using the aggregation function (median in this experiment).



(a) CTSC in 4-cluster setting

(b) STSC in 4-cluster setting

Figure 5.7: Distribution of samples clustered into the blue-encoded cluster using CTSC and STSC. Blue solid lines represent the median of samples, blue dashed lines depict the cluster center, and black dashed lines indicate the samples.

We further examined the subset of samples that were added and removed within the blue-encoded clusters when utilizing CTSC as a substitute for existing approaches. The added samples, represented by blue dash lines in Figure 5.8, are distributed close to their cluster center in Figure 5.7a, and some exhibit an increased trend. Notably, these samples are not captured in the cluster using existing approaches. Conversely, the removed samples, depicted by red dash lines, are concentrated far from the cluster center, representing the samples with unrelated trends which were captured in previous methods. This outcome validates that the proposed scheme CTSC captures fewer unrelated samples than existing STSC approaches. This advantage enhances the interpretability of the results from the aggregation function, allowing it to better represent the properties of clusters. Consequently, the increasing trend in the percentage of the workforce engaged in the occupation of art and culture can only be revealed by CTSC.

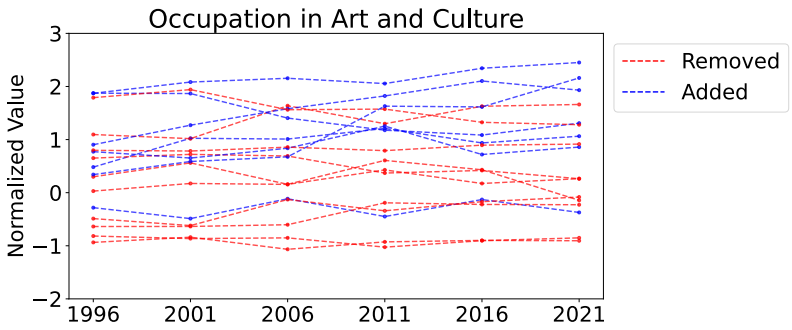
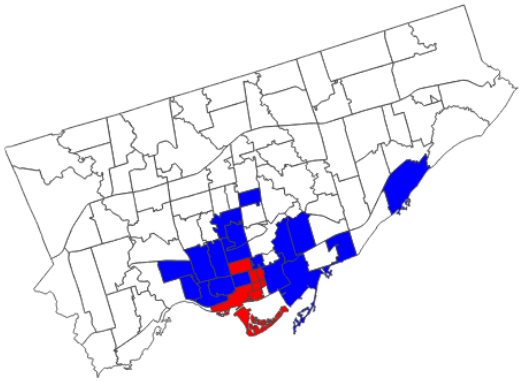


Figure 5.8: The samples added/removed to the blue encoded cluster using CTSC

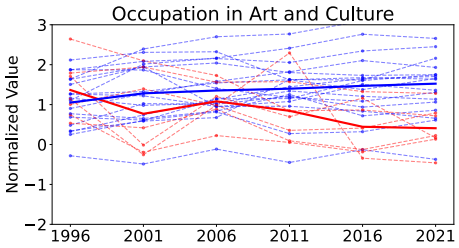
5.4.3 Geological Distribution of Toronto's Workforce in Art and Culture

The clustering results of CTSC, employing the 4-cluster setting, not only unveil the unique increasing trend in the percentage of the workforce engaged in the occupation of art and culture but also maintain the cluster with its samples' median displaying decreasing trends (red solid line in Figure 5.6a), as revealed in existing methods. By combining these two clusters, CTSC illustrates the shift in the geographical distribution of the art and culture workforce in Toronto from 1996 to 2006.

In Figure 5.9a, the FSAs encoded by red choropleths belong to the cluster exhibiting a decreasing trend in overall medians. Most of these FSAs are concentrated in the population center of Downtown Toronto. Conversely, the FSAs encoded by blue choropleths reveal the uniquely discovered increasing trend in their median. These FSAs surround the red choropleths, providing insight into the overall trend of the art and culture workforce moving outside the center of Downtown Toronto.



(a) Geographical distribution of clusters showing a decreasing trend (red) and an increasing trend (blue) in their samples.



(b) Increase trend (blue) and decrease trend (red) shown in the median of samples within each cluster.

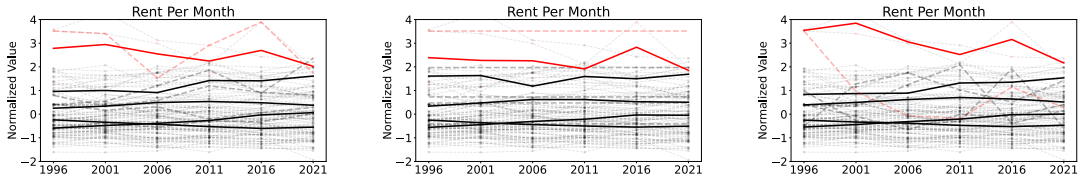
Figure 5.9: The change in geographical distribution of the workforce in art and culture.

Utilizing a scheme from the expanded temporal clustering framework, which permits changes in both clustering center and assignment, effectively reduces the number of unrelated samples in each cluster, thereby minimizing noise in aggregated trends and improving interpretability. The incorporation of this approach has unveiled new trends associated with the geographical distribution of the workforce population, prompting potential investigations into the causes of this migration and justifying the potential utility of this optimal

temporal clustering.

5.5 Trends of Average Monthly Rent in Downtown Toronto

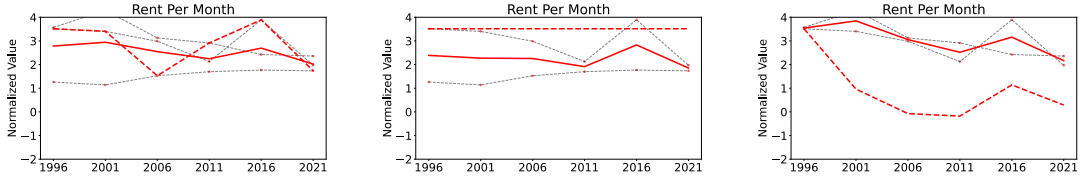
Applying the behavior of the newly proposed schemes, which eliminate unrelated samples to enhance the interpretability of aggregation results, can be extended to various other features. In both STSC (Figure 5.10a) and FUTC (Figure 5.10b) approaches with the 5-cluster setting, the decreasing trend evident in the median of samples in average rent per month for each FSA (depicted by the red solid line) clustered into the red-encoded cluster is not as pronounced as the same trend observed in UTC (Figure 5.10c), the unconstrained variation of CTSC, which allows both cluster center and unbounded cluster assignment changes.



(a) STSC (b) FUTC (c) UTC

Figure 5.10: Clustering results of average rent per month using three different approaches. Red solid lines represent the median of samples in the cluster that exhibits a decreasing trend.

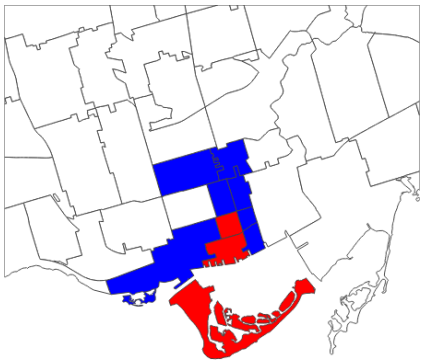
Comparing the samples in red-encoded clusters from three results (Figure 5.11), only UTC completely and accurately captured the only two significantly decreased time series from the temporal data, as shown in Figure 5.11c. The other two approaches either failed to capture all samples with this significant decreasing trend into one cluster (Figure 5.11b) or included samples with no decreasing trend (Figure 5.11a).



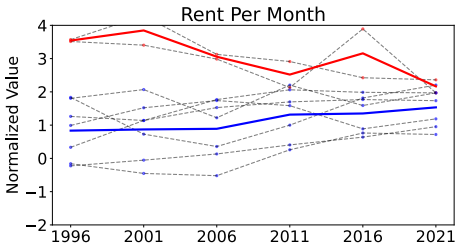
(a) STSC (b) FUTC (c) UTC

Figure 5.11: Samples that were clustered in the red-encoded clusters using three schemes

Although the total sample sizes in these clusters are very small, correctly clustering all samples with decreasing trends can still enhance the interpretability of the cluster meaning and provide unique insights into the Toronto rental market. For instance, in this case, the two FSAs in the UTC result that show significant decreasing trends are 'M5H' and 'M5J', locations in downtown Toronto near Lake Ontario (depicted as red choropleths in Figure 5.12a). However, the blue-encoded cluster, which includes samples adjacent to these two FSAs (depicted as blue choropleths in Figure 5.12a), exhibits an increasing trend in its median aggregation (blue solid line in Figure 5.12b), which is entirely different from its neighbors. Therefore, the use of newly proposed schemes has provided us with a unique insight that cannot be shown using existing methods.



(a) The geo-location of clustered results in Downtown Toronto using UTC with a 5-cluster setting. A group of adjacent FSAs is clustered into two different clusters, indicated by blue and red encodings.)



(b) There are two completed opposite trends shown in samples in these two clusters

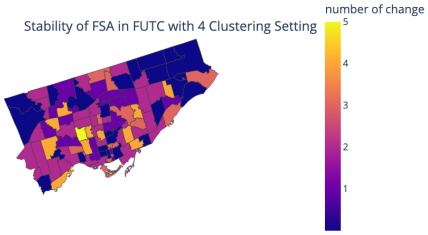
Figure 5.12: Two different trends in rental market shown in center of downtown toronto

5.6 Stability Analysis Using Bounded Label Change

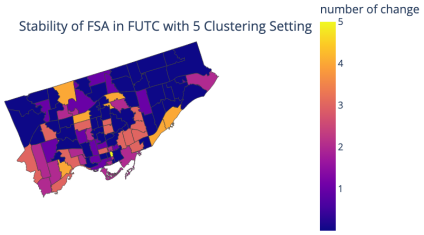
Another enhancement of existing approaches in the proposed temporal clustering framework is the support for the maximum number of allowances in cluster assignment changes. This involves controlling the maximum allowance as a hyperparameter to identify which samples are prone to change from one cluster to another the most, facilitating a more convenient way to analyze the stability of samples across timestamps.

5.6.1 Problem in existing approach

The existing approach of stability analysis, TLA, allows only unbounded label changes and fixed cluster centers throughout all timestamps. As mentioned earlier, to avoid bias caused by different error objectives, we are using Fixed FUTC to simulate the result of TLA in this case study. Figure 5.13 illustrates the number of times each Forward Sortation Area (FSA) changed from one cluster to another, referred to as label change, across all timestamps using FUTC in a 4-cluster setting (Figure 5.13a) or 5-cluster setting (Figure 5.13b). A sample is considered more volatile if it has a higher-than-average number of label changes and stable if it has a lower-than-average count.



(a) 4-Cluster Setting



(b) 5-Cluster Setting

Figure 5.13: Number of times each FSA changed from one cluster to another across all timestamps using FUTC. Light yellow indicates the FSA has the highest number of changes, while deep blue indicates no changes in this FSA.

The primary issue in the results of FUTC is the presence of numerous FSAs that tend to change clusters

when there is no limitation on the maximum number of changes. Moreover, in the unbounded setting, a high number of label changes might indicate that the sample tends to switch between two clusters, but it does not necessarily imply a significant increase or decrease trend across multiple clusters. For instance, in FUTC with a 4-cluster setting (Figure 5.13a), FSA 'M6E' has the highest count of 5 label changes, but its time series in the feature reflecting its label change (Figure 5.14) is relatively stable. Although it might not contradict the definition of volatility in this approach, we are hard to obtain any useful insight from this flat trend.

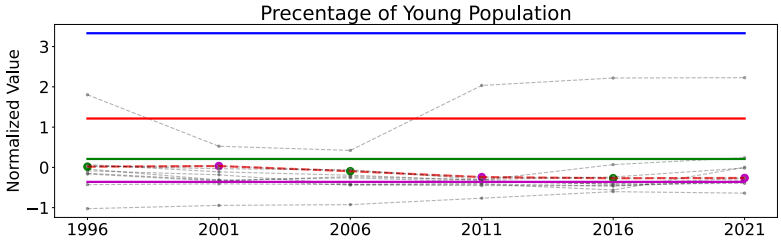


Figure 5.14: The feature 'Total percentage of youth population in FSA' reflects the label change in 'M6E,' shown in the red dashed line. The color of the scatter indicates its class label at the corresponding timestamp, and the solid line represents the fixed cluster center from FUTC's result. The black dashed lines in the background represent 10 randomly selected samples to provide an overview of the overall sample distribution.

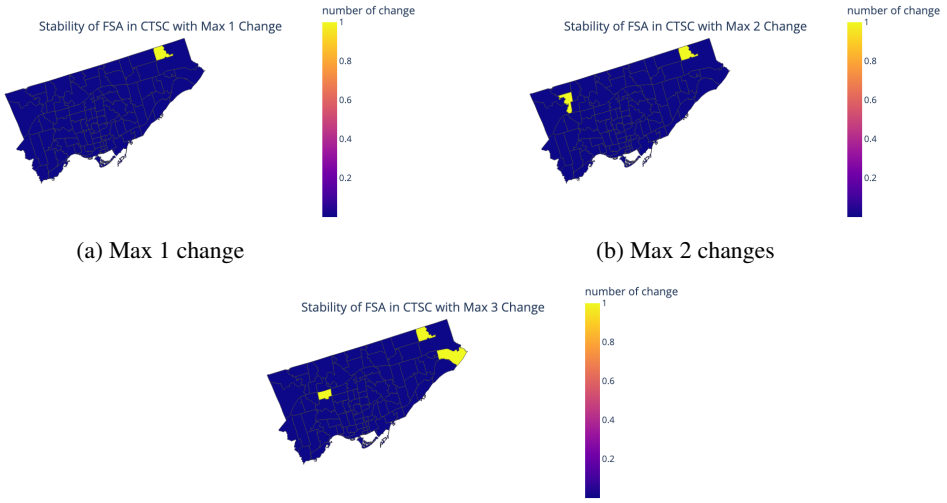
Therefore, relying solely on the total number of changes in samples when using existing approaches with unbounded label changes is not sufficient. This poses a challenge in identifying the specific samples that deviate from the norm, and examining each sample individually becomes infeasible in temporal data with large sample sizes.

5.6.2 Bounded Label Change Approach

To address this issue, the proposed schemes, Fixed Constrained Time Clustering (FCTC) and Constrained Time Series Clustering (CTSC), extend from the existing TLA approach to allow bounded label changes with either fixed or dynamic cluster centers. Figure 5.15 illustrates the number of label changes in Forward Sortation Areas (FSAs) using CTSC with maximum label change allowances of 1 (Figure 5.15a), 2 (Figure 5.15b), and 3 (Figure 5.15c) for each sample. The number of FSAs tends to have a label change significantly drop in these bounded settings, with a positive relationship between the maximum allowance and the total number of FSAs changed (i.e., the higher the maximum allowance, the more the samples changed), which aligns with our expectations. This result allows us to identify the FSAs most prone to change, facilitating the analysis of stability.

In this case, the only FSA that changed its label through all timestamps is 'M1X' in the max 1 change setting, and it consistently exhibits changes in both the max 2 and 3 change settings. Upon investigating its label assignment across all timestamps and features, 'median total income' reflects this change in both the max 1 and 2 change settings. It clearly shows a decreasing trend in Figure 5.16 over time, different from other samples (depicted as black dashed lines), leading to a shift from a cluster with a higher value in its center to another cluster with a lower value. For example, in Figure 5.16a, it transitions from a green-encoded cluster to a red-encoded cluster with a lower value across all timestamps in its center.

'M1X' FSA corresponds to the Upper Rouge area in Scarborough (Figure 5.17), near one of the largest natural conservations, Rouge National Urban Park, in the city of Toronto. A drop in median income in this area can lead to many avenues of study.

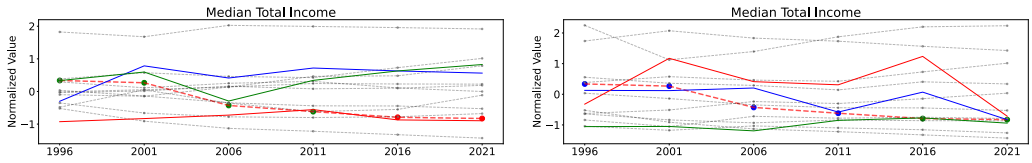


(a) Max 1 change

(b) Max 2 changes

(c) Max 3 changes

Figure 5.15: The number of label changes for each FSA using CTSC with varying maximum allowances for changes per sample. Light yellow indicates the FSA has 1 change, while deep blue signifies it does not have any changes.



(a) Max 1 change

(b) Max 2 change

Figure 5.16: The red dashed line represents the time series of the feature 'Median of Total Income' of 'M1X' reflecting its single label change in both max 1 and 2 change settings. The color-encoded scatter indicates its current label at each timestamp, the solid lines represent cluster centers, and the black dashed lines depict 10 randomly selected samples to provide an overview of the general sample distribution.

More importantly, neither the results of existing approaches (Figure 5.13) show a non-zero label change for 'M1X'. If we solely rely on the unbounded result from the existing TLA approach, this increasing trend will be overlooked in the analysis, even if significant time is spent investigating each sample that changed labels. Therefore, the expansion of existing approaches in stability analysis to allow bounded label changes provides a possibility to extract samples that are most prone to change their label over time within a large group of samples in the unbounded approach. This reduces the analysis workload and unveils undiscovered trends in Toronto census data. However, there are limitations to this bounded change method, such as potentially missing other important but not as significant trends.

5.7 Conclusion

In this case study of Toronto Census data spanning 95 Forward Sortation Areas (FSA) over 25 years, we identified issues in existing approaches that may result in incorrect trends, hard interpretability, or pose challenges for analysis tasks. Subsequently, we introduced and justified a potential solution to address these problems

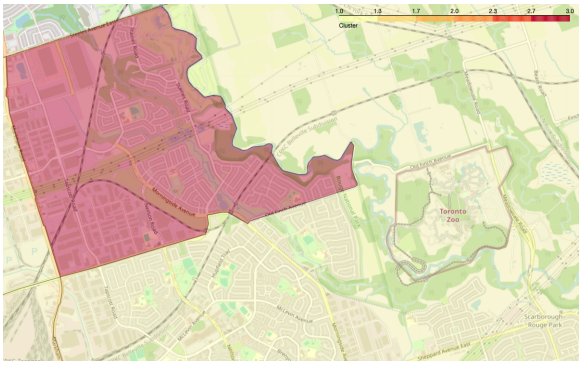


Figure 5.17: Map of 'M1X' showing residences, warehouses, and natural parks within this FSA

through the proposed optimal temporal clustering framework. We provided a rationale for its application and highlighted the benefits demonstrated in a real-world dataset.

Chapter 6

Climate Data

Climate data represents one of the most common types of time series data encountered in our daily lives, making it particularly suitable for temporal clustering analysis. Various locations around the world exhibit distinct climate patterns. For instance, tropical rainforest climates are characterized by high mean temperatures and humidity, while Mediterranean climates display distinct seasonal variations. Clustering the raw climate data into different groups can help researchers reveal the pattern behind climate.

We can also evaluate the trend in climate data against time, which is useful in various fields. Such as weather forecasting, using k-mean clustering with time series air pollution data to predict future days' weather [38]. And sociology study, using global surface temperature across years to reveal the correlation between the history of human civilization and global warming [39].

This chapter will incorporate a case study using real-world historical daily climate data from locations around the world with the proposed optimal clustering schemes. The primary result of this experiment is to apply temporal clustering to climate data to uncover the correlation between the seasonal differences in climate features with its geo-locations and reveal unique climate characteristics of cities around the world. Demonstrating the ability of temporal clustering in climate study.

6.1 Dataset

The dataset is multidimensional climate data obtained from Meteostat [15], comprising daily climate data spanning from 2018 to 2022, encompassing 192 locations worldwide. Among these locations are major cities like Ottawa and Berlin, as well as smaller islands and rural research stations. The dataset encompasses six parameters: maximum, minimum, and mean temperature, wind speed and direction, and atmospheric pressure.

date	country	city	tavg	tmin	tmax	wdir	wspd	pres
21-07-2018	Abkhazia	Sukhumi	23.4	20.9	25.5	329.0	9.3	1009.6
22-07-2018	Abkhazia	Sukhumi	23.5	21.0	25.7	337.0	9.4	1010.0
23-07-2018	Abkhazia	Sukhumi	23.5	21.1	25.5	41.0	8.2	1007.7
24-07-2018	Abkhazia	Sukhumi	24.3	20.8	27.1	10.0	9.3	1004.4
25-07-2018	Abkhazia	Sukhumi	26.5	22.7	30.0	9.0	9.7	1002.0

Table 6.1: First five samples of the original Climate Data from Meteostats

Table 6.1 shows the first 5 samples of the original data. It contains the climate features of Sukhumi, Abkhazia from July 21st to 25th, 2018. In addition, this dataset contains the coordinate information of each location, allowing us to easily identify the pattern in its geo-location.

Figure 6.1 shows the feature distribution among samples. It indicates the presence of outliers and mis-recorded samples in the dataset. For instance, the highest recorded temperature on Earth is 56.7 degrees Celsius. Consequently, it is implausible to have a maximum temperature exceeding 60 degrees Celsius. These outliers have the potential to significantly influence the temporal clustering pattern. As a result, it is recommended to remove outliers that exceed three times the interquartile range (IQR) to ensure a more accurate analysis.

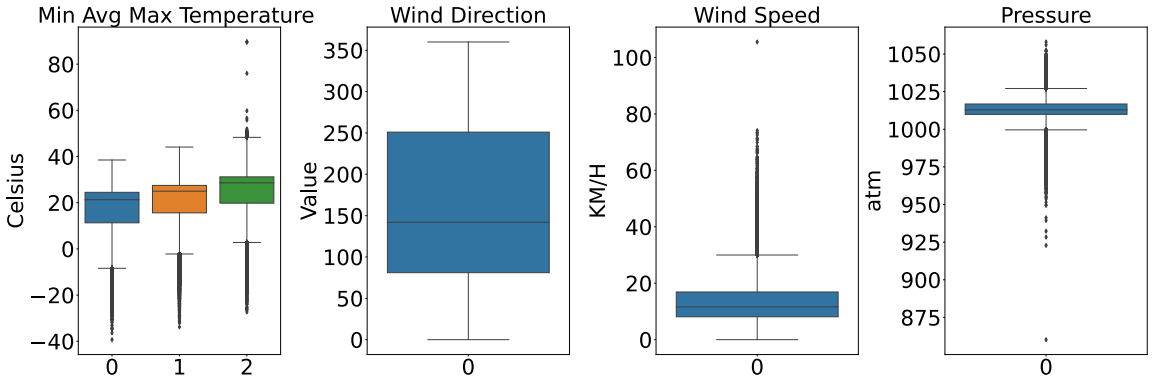


Figure 6.1: Boxplots identified outliers in the features of the dataset

Since the timestamp of samples in the dataset is in days across 5 years, this results in more than 1800 timestamps for each sample. If we treat the daily value of each location as a single time series, the size of the temporal data would be overly complex, which is beyond the current capacity of the proposed optimal schemes. Instead, we aggregated the daily data into the monthly timestamps using the median of each month, and aggregated data of each month across 5 years into a single timestamp. Resulting in a total of 12 timestamps representing the median monthly value across years for each sample.

The dataset also exhibits a maximum of 12.6% missing values. We first removed locations with significantly high percentages of missing values and used linear interpolation to impute the rest missing values to keep the potential trend in the time series. Additional data-cleaning jobs include removing variables with high multicollinearity (minimum and maximum temperature) and standardizing the features into the same scale using mean standardization. The resulting temporal data (Table 6.2 has 72 locations with 12 timestamps and 4 features: average temperature, wind direction, wind speed, and atmospheric pressure.

city	month	avg_tem	wind_dir	wind_spd	pressure
Stanley	05	-1.6	1.3	2.1	-1.0
Saipan	09	1.0	-0.3	0.2	-0.3
Athens	11	-0.5	-0.8	-1.2	1.3
Alofi	04	0.4	-0.2	-0.3	-0.1
Adamstown	09	0.3	-0.3	1.8	1.7

Table 6.2: five samples of the cleaned 5 years monthly (month) climate data from various cities (city) in four features

Figure 6.2 shows the time series in feature average temperature and wind direction of samples. It is hard

to directly observe any pattern in wind direction, which is a great scenario to apply the proposed clustering schemes.

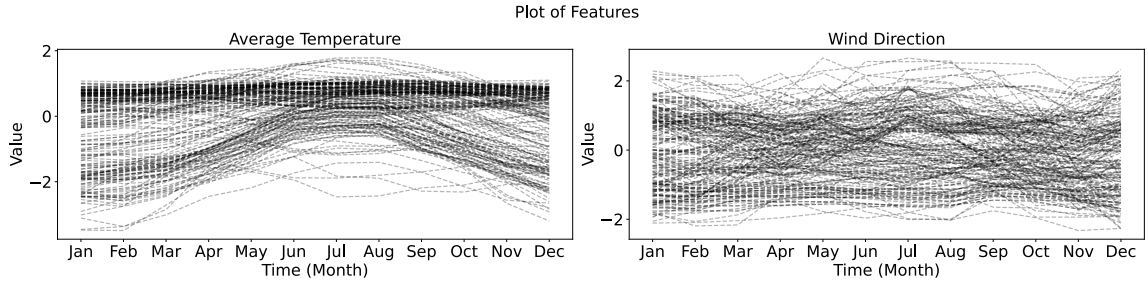


Figure 6.2: Average temperature and Wind direction across 12 months

6.2 Process

In this case study, we employed STSC, the guaranteed optimality variation of the existing Time Series Clustering (TSC) approach. Differing from previous census data, given the limited number of features in this climate temporal dataset (four features), we utilized the k-median objective for MILP. While this approach does not offer the same advantage of runtime reduction as k-center objective, it allows for a fair comparison with existing methods under the same error definition, thus minimizing potential bias.

6.3 Characteristics of Different Clusters

Figure 6.3 illustrates the STSC results with a 3-cluster setting in the 'Average Temperature' feature. The green-encoded cluster consistently exhibits the highest mean temperature throughout the entire year. Closer inspection reveals a relatively high year-round temperature, with no distinct difference between winter (December to February) and summer (June to August). In contrast, both blue-encoded and red-encoded clusters show a seasonal pattern. The blue-encoded clusters exhibit a significant increasing trend from January (winter) to July (summer) and then decrease, while the red-encoded cluster only shows a minor temperature fluctuation throughout the year. The red cluster's lowest monthly temperature is higher than in the blue cluster, and the highest monthly temperature is lower than the blue clusters. Consequently, the effect of seasons on temperature changes in the red cluster's samples is not as significant as in the blue cluster.

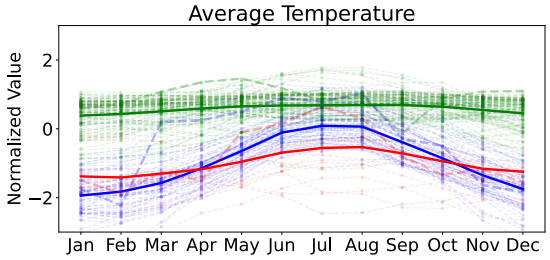
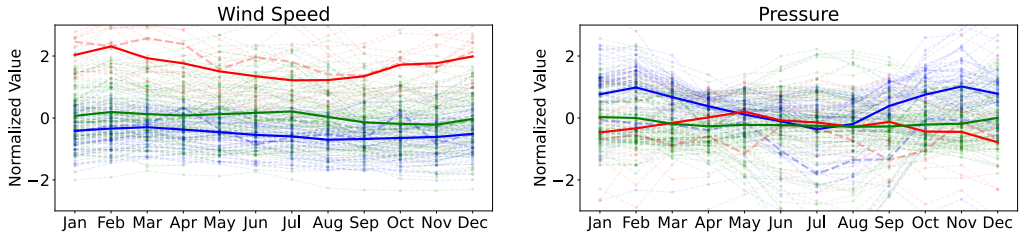


Figure 6.3: Clustering result of STSC in the 'Average Temperature' feature. Solid lines represent mean aggregation of samples in each cluster, and dashed lines denote the cluster centers.

Although the difference in average temperature between the red and blue clusters is relatively small compared to their difference with the green cluster, these two clusters exhibit significant differences in wind speed and pressure characteristics (Figure 6.4). In wind speed (Figure 6.4a), the samples in the red cluster exhibit much higher wind speed than the other two clusters, with a notable drop from June to September. Conversely, the blue cluster demonstrates a relatively stable and low wind speed throughout the years.



(a) Wind Speed

(b) Pressure

Figure 6.4: Clustering result of STSC in features, illustrating the distinct characteristics of red and blue encoded clusters.

However, in the average pressure across months, the trends of the blue cluster are more volatile than the other two clusters, with a similar drop in the Jun-Sep pattern that appears in the wind speed feature of the red cluster. In this case, the red cluster is more stable than the blue cluster. Therefore, using the result of STSC, we can obtain the unique characteristics and trends of three groups of locations in this temporal data. We can also combine this insight with the geological distribution of each group.

6.4 Spatial Analysis of Clustering Result

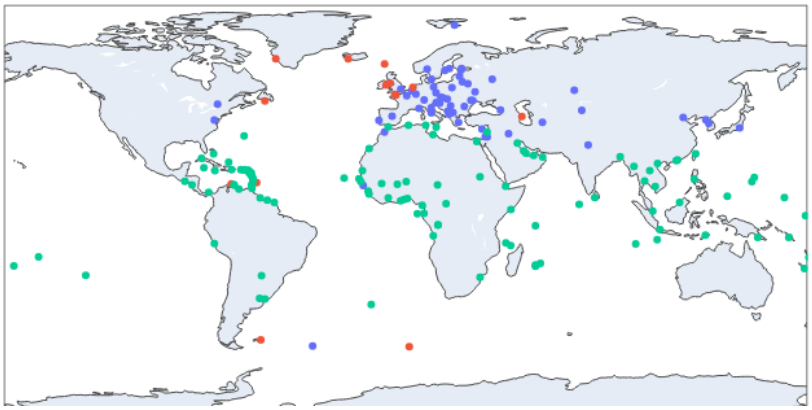


Figure 6.5: Geographical locations of different clusters in STSC.

Using the coordinate information in the given dataset, we mapped samples' cluster assignments with their geographical locations to understand the spatial distribution of different clusters (shown in Figure 6.5).

In green clusters are cities and islands near the equator. Locations in this cluster are situated near the equator, such as Abu Dhabi. These locations also include small islands, corresponding to the green dots in the oceans. Combined with previous findings, it suggests that locations near the equator tend to have consistent temperatures and air pressure throughout the year, and low wind speed. The most important finding is there

are no significant seasonal variations in the locations near the equator. An interesting discovery is that these locations consistently experience eastward wind directions across months.

Samples in the blue cluster contain European Cities and East Asia. Most locations in this cluster are inland cities such as Beijing and Tokyo, represented by the group of blue dots on the map. These locations are clustered within a similar range of longitude. When we consider this spatial distribution alongside the identified climate patterns, which encompass distinct seasonal variations and consistently low wind speeds throughout the year, a compelling pattern emerges. Given that most of the locations in this cluster are major populated cities situated within these two continents, it suggests that the patterns we've uncovered for this cluster can be indicative of common climate characteristics shared by major populated cities.

The red cluster is the fewest samples among all three clusters, but all locations in this cluster are coastal locations around the globe, without a clear longitude characteristic. Combining with clustering results in different features, the locations in this cluster tend to have relatively high wind speed, a distinct but not significant seasonal pattern in average temperature, and relatively stable but low pressure across months.

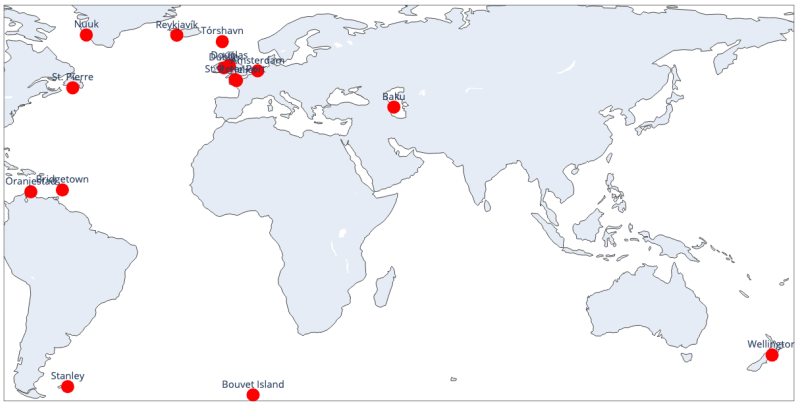


Figure 6.6: Coastal locations in the red cluster.

More importantly, we cannot observe this difference using the existing TSC approach (Figure 6.7). This unique cluster, with all coastal cities, can bring valuable insight to climate study.

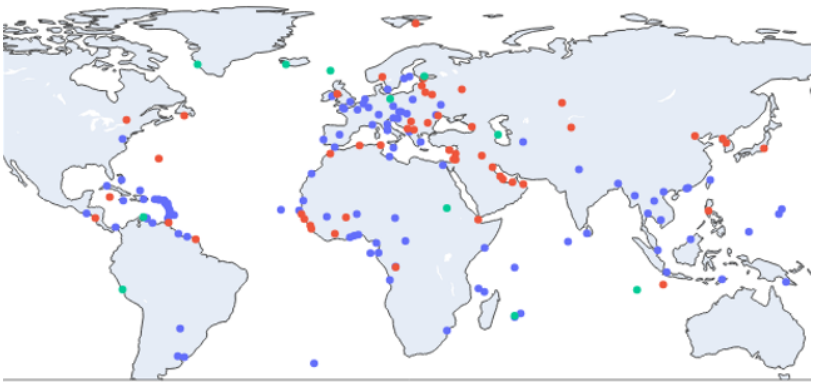


Figure 6.7: Geographical locations of different clusters in TSC, the mentioned spatial pattern is not revealed in this approach.

6.5 Conclusion

In this case study, we applied STSC, the guaranteed optimality variation of the Time Series Clustering (TSC), to real-world climate temporal data, revealing a unique coastal cities cluster not present in the clustering result of the general TSC approach with the same error definition. It showcases a real-world difference between non-optimal and optimal results and the possible insight we can obtain from these differences. Therefore, this case study validates the potential usefulness of optimality in this temporal clustering framework.

Chapter 7

Conclusion

In this study, we undertook a comprehensive review of the temporal clustering framework, aiming to address the gaps present in existing methodologies within the design space. The focus encompassed both dynamic and fixed cluster centers and assignments, with a key emphasis on ensuring the optimality of the clustering results. Through a meticulous comparison of performance using synthetic data, we evaluated existing approaches alongside different objective and error definitions in the clustering context.

The results demonstrated that our proposed framework, featuring dynamic centers but fixed assignments, outperforms established methods such as Time Series clustering and Dynamic Time Wrapping. Notably, it excels in reducing the maximum distance between cluster samples and their centers while maintaining a high level of purity. Additionally, the incorporation of diverse objectives, as introduced by Tosanwumi to enhance efficiency, exhibited a substantial reduction in runtime during our evaluation, affirming the effectiveness of our proposed approach.

This project further contributed by introducing two insightful case studies. Leveraging Census data from the Toronto Forward Sortation Area spanning 1996 to 2021 and historical climate data from 192 locations worldwide, we applied our optimal temporal clustering framework. The results from both case studies unveiled unique insights that were either absent or misinterpreted in existing approaches, such as discerning shifts in the distribution of the art and culture workforce in Toronto. Beyond their contribution to potential future studies in this domain, these case studies also offered practical applications and usage ideas for our newly formulated frameworks.

Nevertheless, the current framework is not without its limitations. One notable constraint lies in the substantial increase in the number of constraints in Mixed-Integer Linear Programming (MILP) formulations as the total number of time series or timestamps within each time series grows. This escalation results in infeasible runtimes for the optimization problem, thereby restricting the framework's applicability to real-world studies with larger feature sets or sample sizes. Consequently, our case studies suggest avenues for potential improvements in the scalability of the proposed framework to accommodate the growing trend of sample sizes in contemporary studies.

Bibliography

[1] M. Chiş, S. Banerjee, and A. E. Hassanien, "Clustering time series data: an evolutionary approach," *Foundations of Computational, IntelligenceVolume 6: Data Mining*, pp. 193–207, 2009.

[2] P. Esling and C. Agon, "Time-series data mining," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 1–34, Dec. 2012.

[3] T.-c. Fu, "A review on time series data mining," *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 164–181, Feb. 2011.

[4] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information systems*, vol. 53, pp. 16–38, 2015.

[5] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[6] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[7] T. K. Dey, A. Rossi, and A. Sidiropoulos, "Temporal clustering," *arXiv preprint arXiv:1704.05964*, 2017.

[8] M. Hoai and F. De la Torre, "Maximum margin temporal clustering," in *Artificial Intelligence and Statistics*. PMLR, 2012, pp. 520–528.

[9] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320305001305

[10] E. C. Delmelle, "Mapping the dna of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change," *Annals of the American Association of Geographers*, vol. 106, no. 1, pp. 36–56, 2016.

[11] F. Dias and D. Silver, "Neighborhood dynamics with unharmonized longitudinal data," *Geographical Analysis*, vol. 53, no. 2, pp. 170–191, 2021.

[12] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?" *Pattern Recognition*, vol. 93, pp. 95–112, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319301608

[13] A. Chembu, S. Sanner, H. Khurram, and A. Kumar, "Scalable and globally optimal generalized l1 k-center clustering via constraint generation in mixed integer linear programming," 2023.

[14] "Dynamic Time Warping," in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer, 2007, pp. 69–84.

[15] C. S. Lamprecht, "Meteostat python." [Online]. Available: https://orcid.org/0000-0003-3301-2852

[16] "CHASS Data Centre," Aug. 2023, [Online; accessed 11. Dec. 2023]. [Online]. Available: https://dc1.chass.utoronto.ca

[17] N. S. Madiraju, "Deep temporal clustering: Fully unsupervised learning of time-domain features," Ph.D. dissertation, Arizona State University, 2018.

[18] E. J. Keogh and M. J. Pazzani, "A simple dimensionality reduction technique for fast similarity search in large time series databases," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2000, pp. 122–133.

[19] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Locally adaptive dimensionality reduction for indexing large time series databases," in *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, 2001, pp. 151–162.

[20] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," *ACM Transactions on Database Systems (TODS)*, vol. 27, no. 2, pp. 188–228, 2002.

[21] C. Ratanamahatana, E. Keogh, A. J. Bagnall, and S. Lonardi, "A novel bit level time series representation with implication of similarity search and clustering," in *Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005. Proceedings 9*. Springer, 2005, pp. 771–777.

[22] A. Bagnall and G. Janacek, "Clustering time series with clipped data," *Machine learning*, vol. 58, pp. 151–178, 2005.

[23] S. Chu, E. Keogh, D. Hart, and M. Pazzani, "Iterative deepening dynamic time warping for time series," in *Proceedings of the 2002 SIAM International Conference on Data Mining*. SIAM, 2002, pp. 195–212.

[24] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*. Springer, 2001, pp. 420–434.

[25] J. Mennis and D. Guo, "Spatial data mining and geographic knowledge discovery—an introduction," *Computers, Environment and Urban Systems*, vol. 33, no. 6, pp. 403–408, 2009.

[26] J. R. Logan and W. Zhang, "Identifying ethnic neighborhoods with census data: Group concentration and spatial clustering," *Spatially integrated social science*, pp. 113–126, 2004.

[27] F. Bação, V. Lobo, and M. Painho, "Clustering census data: comparing the performance of self-organising maps and k-means algorithms," in *KDNet Symposium, Bonn, Germany*, 2004.

[28] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," *Data Mining and Knowledge Discovery*, vol. 25, pp. 478–510, 2012.

[29] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[30] G. Marti, P. Very, P. Donnat, and F. Nielsen, "A Proposal of a Methodological Framework with Experimental Guidelines to Investigate Clustering Stability on Financial Time Series," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 09–11.

[31] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, "Synthetic Data – what, why and how?" *arXiv*, May 2022.

[32] M. L. Fisher, "An Applications Oriented Guide to Lagrangian Relaxation," *Interfaces*, Apr. 1985. [Online]. Available: https://pubsonline.informs.org/doi/abs/10.1287/inte.15.2.10

[33] G. H. Knibbs, "The Analysis of a Census," *Quarterly publications of the American Statistical Association*, Jun. 1920. [Online]. Available: https://www.tandfonline.com/doi/pdf/10.1080/15225445.1920.10503784?casa_token=iZc_oKQIfioAAAAA:27WBSLaLmtFbJIr_WUsffYwCSkyrYxL9m73ij92KIKcxSE-0CTVOGvjiC1i2V3mOPYxRmVssrLul4w

[34] K. A. Woodrow-Lafield, "Census analysis," in *Routledge International Handbook of Migration Studies*. Routledge, 2019, pp. 539–552.

[35] "Census of canada: Profile data for toronto at the forward sortation areas (fsa)," in *Canadian Census Analyser*. Toronto, Ontario: Statistic Canada, 2021. [Online]. Available: http://dc.chass.utoronto.ca.myaccess.library.utoronto.ca/cgi-bin/census/2021/displayCensus.cgi?year=2021&geo=fsa

[36] E. C. Budd and D. F. Seiders, "The impact of inflation on the distribution of income and wealth," *The American Economic Review*, vol. 61, no. 2, pp. 128–138, 1971. [Online]. Available: http://www.jstor.org/stable/1816985

[37] L. G. Svensson, "Occupations and Professionalism in Art and Culture," *P&P*, vol. 5, no. 2, Aug. 2015.

[38] S. Chakraborty, N. K. Nagwani, and L. Dey, "Weather Forecasting using Incremental K-means Clustering," *arXiv*, Jun. 2014.

[39] M. Mudelsee, "Trend analysis of climate time series: A review of methods," *Earth-Sci. Rev.*, vol. 190, pp. 310–322, Mar. 2019.