A Time-Series Clustering Algorithm for Analyzing the Changes of Mobility Pattern Caused by COVID-19

Ziyi Zhang
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX, US
zyzhang1996@tamu.edu

Diya Li
Department of Geography
College of Geosciences
Texas A&M University
College Station, TX, US
diya.li@tamu.edu

Zhe Zhang
Department of Geography
College of Geosciences
Texas A&M University
College Station, TX, US
zhezhang@tamu.edu

Nicholas Duffield
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX, US
duffieldng@tamu.edu

ABSTRACT

The coronavirus (COVID-19) has spread to more than 135 countries and continues to spread. The virus sickened more than 90,201,652 people until January 2021 and caused 1,937,091 deaths in the world. So far, social distancing plays a vital role in controlling the coronavirus. Governments issued restrictions on traveling, institutions cancel gatherings, and citizens socially distance themselves to limit the spread of the virus. This paper aims to develop a novel time-series clustering algorithm to analyze the changes in mobility patterns caused by the COVID-19. This work will produce broader impacts in many areas, such as helping local governments locate the medical facilities and improving the social distancing recommendations for infectious disease control.

CCS CONCEPTS

• **Information systems** → Information systems applications; Spatial-temporal systems; Data Mining; • **Computing methodologies** → Models of computation; Timed and hybrid models;

KEYWORDS

Time-series clustering; COVID-19; mobility pattern

ACM Reference format:

Ziyi Zhang, Diya Li, Zhe Zhang, and Nicholas Duffield. 2021. A Time-Series Clustering Algorithm for Analyzing the Changes of Mobility Pattern Caused by COVID-19. In *1st ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility (HANIMOB'21), November 2, 2021, Beijing, China. ACM, New York, NY, USA, 4 pages.* https://doi.org/10.1145/3486637.3489489

HANIMOB'21, November 2, 2021, Beijing, China

ACM ISBN 978-1-4503-9122-1/21/11□$15.00
https://doi.org/10.1145/3486637.3489489

1 INTRODUCTION

The coronavirus (COVID-19) has caused a huge negative influence on the global. The US government has issued restrictions on traveling, the institution has canceled gatherings, and the citizens have socially distanced themselves to limit the increase of confirmed cases. COVID-19 has dragged down the global economy by forcing humans to practice social distancing and restricting business travels. The review of the literature indicated that Geographical Information Systems have been used as an important tool to measure risk-informed situation awareness to support spatial decision-making during the pandemic [1]. For example, authors in [2] modeled human dynamics based on social media and urban infrastructure data to support spatial decision making, and the authors in [3] developed an interactive web-based platform that integrates data streams with visual elements such as maps and time-series plots to represent the mobility patterns in response to COVID-19. Authors in [4] indicated that the beneficial effects of social-distancing measures are substantial by modeling the relationship between transmission and mobility patterns.

With the development of the Internet of Things (IoT) and sensor technology, more and more time-series data are available for research purposes [5]. For example, SafeGraph [6] data provides a reliable point of interest, footprints, and foot traffic data, and the Bureau of Transportation Statistics (BTS) provides daily travel data [7] during the COVID-19 public health emergency.

Extracting and analyzing patterns from time series data is important since we can learn some useful knowledge that is hidden in the data [8]. Clustering analysis has been widely used in spatiotemporal data mining to group similar objects to identify the hidden spatial pattern [9]. Unsupervised time-series clustering is the process of classifying time series when there is no early knowledge about classes [5].

In general, there are three common ways to cluster time series data: 1) *Whole-time series clustering* conduct clustering on the raw time series without doing any transformation on original time series; 2) *Subsequence clustering* means clustering the subsequences generated from original time series; 3) *Point clustering* is based on the similarity of time-point values [5]. Additionally, time-series clustering can be divided into 1) a

Shape-based approach, the time series clustering algorithm clusters the time series data according to the similarities of the data; 2) a *Feature-based* approach, we need to extract features from the original time series and then apply K-means, DBSCAN, or some clustering algorithms on these features [10].

Recent studies have explored the COVID-19 related data by applying time-series clustering. For example, the authors of [11] applied K-means on home dwell records data derived from SafeGraph [6], which reflect the length of time people stay at home, to generate spatiotemporal patterns of mobility in the Metro Atlanta area, and some demographic factors such as economic status, race, and education were integrated to explain the disparity of the generated results.

However, we found only apply K-means to cluster time series may generate a misleading result if there is a time lag between two-time series. For example, time-series 2 is the same as time-series 1 but with a time lag equals to 1s illustrated in Figure 1, if we use the Euclidean distance in the K-means algorithm to measure the similarity of two-time series, time lag enables the similarity between two-time series to be 9 by calculating the distance between each pair of points rather than 0. Thus, we proposed to combine the K-means algorithm and Dynamic Time Warping (DTW) to match two-time series since DTW can eliminate the time lags between two different time series, and it has been proved to work well in time-series tasks [12]. We will present how DTW works in Section 3.2.

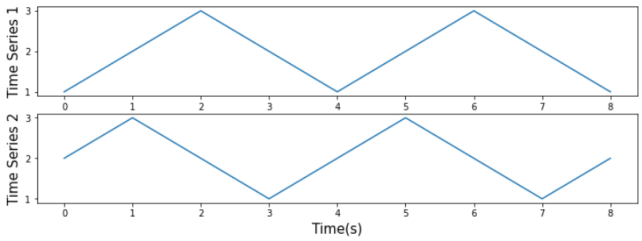

Figure 1: Illustration of the time lag (1s) between two-time series.

2 DATASET

The mobility data used in our research is derived from the *Bureau of Transportation Statistics (BTS)* produced by the Maryland Transportation Institute and Center for Advanced Transportation Technology Lab at the University of Maryland [7]. The "trips" in the dataset are defined as movements that include staying longer than 10 minutes at an anonymized location away from home. The types of transportation include driving, rail, and air transit. The daily travel estimates are merged from multiple data sources, which only contain the mobile devices that meet the quality standards to ensure data quality and consistency.

The second dataset used in our work is *Cylinder-Bell-Funnel (CBF)* dataset, which is a simulated artificial dataset explained in the work [13]. There are three different classes in this dataset: cylinder, bell, and funnel, we extracted some samples from the *CBF* dataset and visualized them in Figure 2, Figure 3, and Figure 4. The data from each class are standard normal noise plus an offset term that differs for the class. The dataset includes three classes and 930-time series, and each time series contains 128-time steps. The *CBF* dataset was used in our work to evaluate the performance of different methods.

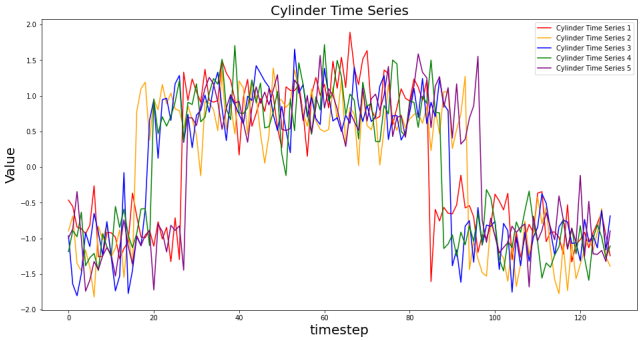

Figure 2: Visualization of Cylinder time-series samples in CBF.

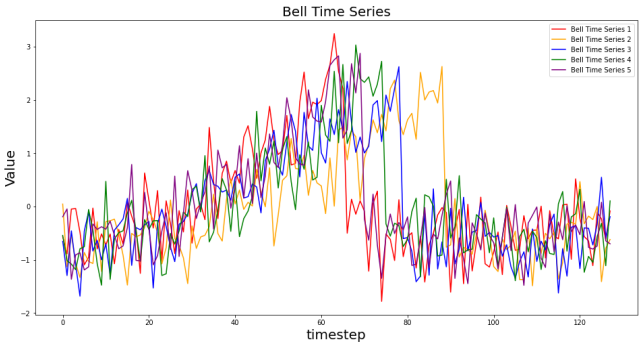

Figure 3: Visualization of Bell time-series samples in CBF.

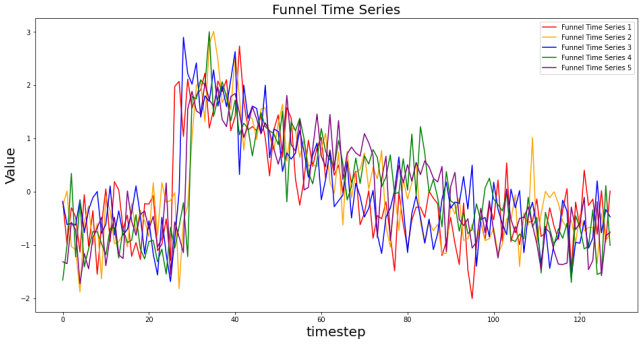

Figure 4: Visualization of Funnel time-series samples in CBF.

3 METHODOLOGY

3.1 Data Preprocessing

The mobility data introduced in Section 2 contains the number of trips in n different mileage categories. To better compare the variations among different counties, we first calculated the average weighted distance using the equation below:

$$D_i = \sum_{j=1}^{n} w_{ij} N_{ij} \quad (1)$$

where D_i is the average distance (meter) at a specific date at location i, N_{ij} is the number of trips in the j^{th} mileage category at location i, and w_{ij} is the weight of the j^{th} mileage category at location i. We set w_{ij} to be the median of the j^{th} mileage category at location i in our work. Next, we used linear interpolation to fill in the missing values and merged daily data into time-series for each month. We chose to add a 7-day moving average to each time series to reduce the impact of certain outliers and noise (See Figure 5). The 7-day moving average for $t = i$ is defined as:

$$MA[i] = \frac{\sum_{t=i-3}^{i+3} T_t}{7} \quad (2)$$

where T_t is the value of the original time series at time t.

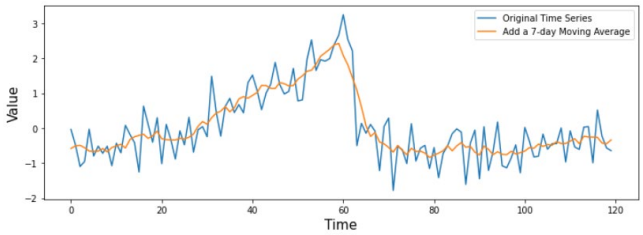

Figure 5: Time-series after adding a 7-day moving average.

We selected the time-series data during the months' March, April, and May 2020, since the social-distancing measures and COVID-19 cases changed significantly in some states in the US during these three months. After that, we used the selected moving average time-series data as the input for the time-series clustering analysis.

3.2 DTW-based K-means Algorithm

The algorithm that combines K-means with dynamic time warping (DTW-based K-means) is described in Figure 6. Applying K-means with dynamic time warping as a metric to cluster time-series can be detailed as: for a dataset T which contains n time series, 1) Initialize K central time series from dataset T randomly; 2) Calculate the dynamic time warping between each time series in T and each current K central time series; 3) Assign the time series to the central time series whose dynamic time warping from the central time series minimum of all central time series; 4) Recalculate the new K central time series; 5) Recalculate the dynamic time warping between each time series and each new obtained central time series; 6) Stop until the time series will not be reassigned, otherwise repeat from Step 2. Dynamic time warping mentioned in the above time series clustering algorithm is described as 1) Given two-time series $P = P_1, P_2 \dots, P_n$ and $Q = Q_1, Q_2 \dots, Q_n$; 2) Create an $n \times n$ matrix whose i, j^{th} element is the Euclidean distance between p^i and q^j, the objective of dynamic time warping is to find the minimum cumulative distance through the matrix; 3) Define the path through the matrix be $M = m_1, m_2, \dots, m_k$; 4) Dynamic time warping is the minimum Euclidean distance: $M^* = argmin_M(\sqrt{\sum_{m=1}^{K} m_k})$.

Algorithm 1 K-means for Time Series Clustering

1: **procedure** INPUT(T, K) ▷ N-size Time Series Dataset: T_N; Number of Cluster: K
2: Randomly initialize K central time series $t = t_1, t_2, ..., t_K$.
3: Calculate DTW between each time series in $T = T_1, T_2, ..., T_N$ and each central time series in $t = t_1, t_2, ..., t_K$.
4: Assign the time series in T to the central time series t_i whose DTW from the central time series t_i is minimum of all K central time series, and finally forms K cluster: $C = C_1, C_2, ..., C_k$.
5: Recalculate the new K central time series in $C = C_1, C_2, ..., C_k$.
6: Recalculate the DTW between each time series in T and new obtained K central time series.
7: Stop until C will not change, otherwise repeat from Step 3.
8: **return** C

Algorithm 2 Dynamic Time Warping (DTW)

1: **procedure** INPUT(P, Q) ▷ Time Series P with length n; Time Series Q with length n
2: **for** $i = 1$ to n **do**
3: **for** $j = 1$ to n **do**
4: $Matrix(i, j) \leftarrow EuclideanDistance(p_i, q_j)$
5: Define a path through the matrix:$M = m_1, m_2, ..., m_k$
6: $DTW \leftarrow argmin_M(\sqrt{\sum_{m=1}^{K} m_k})$
7: **return** DTW

Figure 6: Algorithms of K-means for times-series clustering and dynamic time warping.

3.3 Evaluation Metric

In this project, we apply the elbow method to determine the optimal numbers of clusters $K = 4$. To make the generated clusters interpretable, we defined the number of mobility levels according to $K = 4$, and used the inner average distance (meter) as the representation for each cluster, which can be expressed as:

$$R_{Cluster} = \frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{T} V_{ij} \quad (3)$$

where L is the number of locations, T is the number of time points for each county, and V is the average distance of the j^{th} time point at location i. According to the value of $R_{Cluster}$ and optimal numbers clusters $K = 4$, we defined the K mobility levels from low to high. The result of representations for each cluster in March, April, and May are listed in Table 1.

	Low	ML	MH	High
March	1611.73	1722.59	1834.79	4601.95
April	1310.01	1543.28	2320.67	4187.94
May	1503.76	1627.11	2255.92	3059.13

Table 1: Representations (meter) for each cluster in March, April, and May.

4 RESULTS

4.1 Performance Evaluation

We first compared the performance of the K-means algorithm and DTW-based K-means on the *CBF* datasets. We found that DTW-

based K-means produced more reliable results (e.g., with a better Rand Index = 0.8692) than K-means (Table 2). Rand Index is used in the work [10] to validate the performance of proposed methods on time-series clustering task, which can be expressed as:

$$RI = \frac{TP + TN}{n(n-1)/2} \quad (4)$$

where the true positives (TP) are the number of pairs of time series that are correctly put in the same cluster, the true negatives (TN) are the number of pairs that are correctly put in different clusters and n is the size of the dataset.

However, the running time of K-means is shorter than DTW-based K-means. Here, we used DTW-based K-means to generate more reliable cluster maps to illustrate the distribution of mobility changes under COVID-19.

Method	Rand Index	Time (S)
K-means	0.6995	**5.82**
DTW-K-means	**0.8692**	190.35

Table 2: Illustration of the algorithm performance.

4.2 Mobility Pattern Analysis

We defined four mobility levels (low, medium-low, medium-high, and high) based on $R_{Cluster}$ in our cluster analysis. The spatial distributions of the clusters in March, April, and May are shown from Figure 7 to Figure 9. The lighter the color represents the higher mobility changes of the cluster. In this case, the yellow, green, blue, and dark purple represent the high, medium-high, medium-low, and low mobility change categories. Results indicated that April met a decrease when compared to March, from the spatial distribution (Figure 7, Figure 8), representation for each cluster in each month (Table 1), and number in each cluster (Table 3), which may be caused by the announcement of social-distancing policies.

	March	April	May
Low	1487	2384	2017
ML	1015	537	819
MH	532	84	104
High	103	132	197

Table 3: Number of counties in each cluster

Besides, we can see that the medium-high clusters occupied most counties in the west and mid-west of the US in March. The situation significantly improved for the western regions of the US in April, especially for the states Nevada, Arizona, Wyoming, and Colorado. Most of the counties in those states moved into low and medium-low categories. However, some states in the mid-west of the US remained in the medium-high category. Some of the counties in the mountain states remained in the high cluster category from April to May.

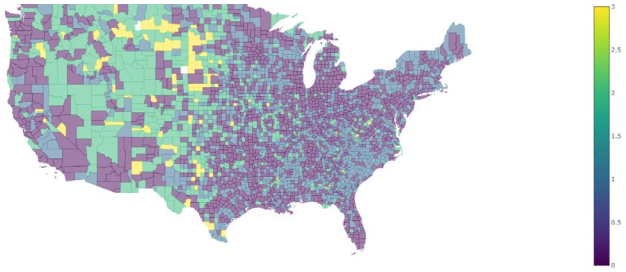

Figure 7: Illustration of March mobility time-series clusters.

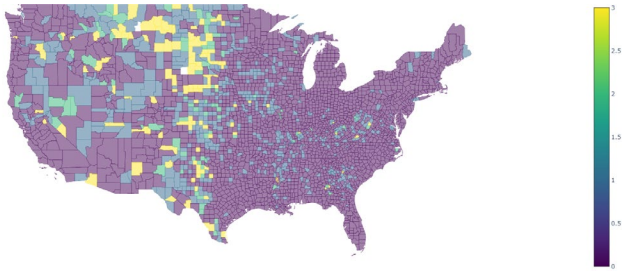

Figure 8: Illustration of April mobility time-series clusters.

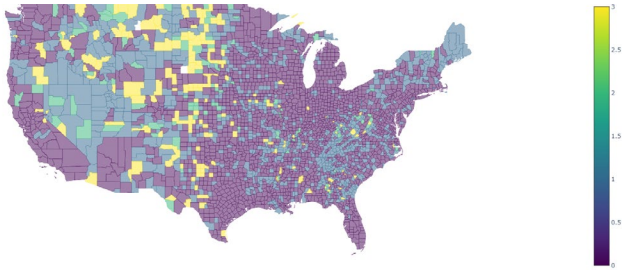

Figure 9: Illustration of May mobility time-series clusters.

5 CONCLUSION

In this paper, we developed a novel DTW-based K-means algorithm to analyze time-series mobility data to observe the mobility change caused by the COVID-19 pandemic. Results indicated that our DTW-based K-Means algorithm has produced better performance than the K-means algorithm. In the future, we are going to improve the computational efficiency and meanwhile keep the reliability of the results from two aspects: 1) By incorporating a fast-DTW [14] algorithm into K-means to lower the running time; 2) Considering the differences in patterns between weekdays and weekends, it is necessary to conduct feature selection of the input time-series firstly when performing a DTW-based K-means algorithm. DTW-based K-means algorithm can also be applied to cluster animal mobility time-series data to generate more reliable spatial patterns, which is helpful for people to formulate more effective animal protection strategies.

ACKNOWLEDGMENTS

This work is supported by the Texas A&M Institute of Data Science Data Resource Development Program.

REFERENCES

[1] Li, D., Chaudhary, H., & Zhang, Z. (2020). Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. International Journal of Environmental Research and Public Health, 17(14), 1–23. https://doi.org/10.3390/ijerph17144988

[2] Zhang, Z., Yin, D., Virrantaus, K., Ye, X., & Wang, S. (2021). Modeling human activity dynamics: an object-class oriented space–time composite model based on social media and urban infrastructure data. Computational Urban Science 2021 1:1, 1(1), 1–13. https://doi.org/10.1007/S43762-021-00006-X

[3] Gao, S., Rao, J., Kang, Y., Liang, Y., & Kruse, J. (2020). Mapping county-level mobility pattern changes in the United States in response to COVID-19. SIGSPATIAL Special, 12(1), 16–26. https://doi.org/10.1145/3404820.3404824

[4] Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E. C., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N. F., Cattarino, L., Cooper, L. V., Coupland, H., Cucunuba, Z. M., Cuomo-Dannenburg, G., Dighe, A., Djaafara, B. A., Dorigatti, I., Eales, O. D., van Elsland, S. L., Nascimento, F. F., … Donnelly, C. A. (2021). Reduction in mobility and COVID-19 transmission. Nature Communications, 12(1), 1–9. https://doi.org/10.1038/s41467-021-21358-2

[5] Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. Information Systems, 53, 16–38. https://doi.org/10.1016/j.is.2015.04.007

[6] Places Data & Foot Traffic Insights | SafeGraph. (n.d.). Retrieved October 6, 2021, from https://www.safegraph.com

[7] Trips by Distance | Open Data | Socrata. (n.d.). Retrieved May 27, 2021, from https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv

[8] Han, J., Kamber, M., and Pei, J., 2012. Data mining concepts and techniques. 3rd ed. Waltham, MA: Morgan Kaufman, MIT Press. https://tinman.cs.gsu.edu/~zcai/course/47406740/Slides/Chapter%201%20Introduction%20to%20Data%20Mining.pdf

[9] P. Rai, S. Singh, A survey of clustering techniques, Int. J. Comput. Appl. 7 (12) (2010)1–5. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.5219&rep=rep1&type=pdf

[10] Ma, Q., Zheng, J., Li, S., & Cottrell, G. W. (2019). Learning Representations for Time Series Clustering. Advances in Neural Information Processing Systems, 32.

[11] Huang, X., Li, Z., Lu, J., Wang, S., Wei, H., & Chen, B. (2020). Time-series clustering for home dwell time during COVID-19: What can we learn from it? ISPRS International Journal of Geo-Information, 9(11), 675. https://doi.org/10.3390/ijgi9110675

[12] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. 2. https://arxiv.org/abs/1003.4083v1

[13] Saito, N. (2000). LOCAL FEATURE EXTRACTION AND ITS APPLICATIONS USING A LIBRARY OF BASES. In Topics in Analysis and Its Applications (pp. 269–451). WORLD SCIENTIFIC. https://doi.org/10.1142/9789812813305_0005

[14] AssentIra, WichterichMarc, KriegerRalph, KremerHardy, & SeidlThomas. (2009). Anticipatory DTW for efficient similarity search in time series databases. Proceedings of the VLDB Endowment, 2(1), 826–837. https://doi.org/10.14778/1687627.1687721