# Comparison of geospatial trajectory clustering and feature trajectory clustering for public transportation trip data

Hector Cogollos Adrian[1][0000−0002−6718−1008], Bruno Baruque Zanon[1][0000−0002−4993−204X], Santiago Porras Alfonso[2][0000−0003−4331−3085], and Petr Dolezel[3][0000−0001−6194−0467]

[1] Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Digitalización, Escuela Politécnica Superior, Universidad de Burgos, Av. Cantabria s/n, 09006, Burgos, Spain.
[2] Metaheurísticos (GRINUBUMET), Departamento de Economía Aplicada, Facultad de Ciencias Económicas y Empresariales, Universidad de Burgos, Pza. de la Infanta Dª. Elena, s/n, 09001, Burgos, Spain.
[3] University of Pardubice, Faculty of Electrical Engineering and Informatics, Studentska 95, 532 10 Pardubice, Czech Republic

**Abstract.** One of the techniques for the analysis of travel patterns on a public transport network is the clustering of the users movements, in order to identify common patterns. This paper analyses and compares two different methodologies for public transport trajectory clustering: feature clustering and geospatial trajectory clustering. The results of clustering trip features, such as origin, destination, or distance, are compared against the clustering of travelled trajectories by their geospatial characteristics. Algorithms based on density and hierarchical clustering are compared for both methodologies. In geospatial clustering, different metrics to measure distances between trajectories are included in the comparison. Results are evaluated by analysing their quality through the silhouette coefficient and graphical representations of the clusters on the map. The results show that geospatial trajectory clustering offers better quality than that obtained through feature clustering. Also, in the case of long and complete trajectories, density clustering using ERP distance outperforms other combinations.

**Keywords:** HDBSCAN · Agglomerative Clustering · ERP · DTW

## 1 Introduction

An increasing interest in promoting less polluting means of transportation, such as public transportation, has been apparent in recent years for most developed countries citizens. This change has been encouraged by institutions like the United Nations through the Sustainable Development Goals or the European Union through the European Green Deal. To incentivize the use of public transportation, the understanding of the users behaviours and habits is considered of capital importance, so the services can adapt better to their needs [1, 2].

In order to identify patterns that are relevant for a significant number of users, a massive number of trips need to be analysed. To cope with this amount of data, a straightforward approach is to cluster their registered movements in an automated way and then analyse the characteristics of those said clusters.

Two different methodologies for clustering public transport data are considered in this study. The first methodology focuses on extracting a certain number of features from each trip, in order to perform clustering according to those features, while the second methodology uses ordered sequences of GPS coordinates to measure similarity between trips in order to cluster them.

The objective of this work is to expand the area of knowledge in this application field and to try to determine which methodology offers more reliable results. To make this comparison more complete and understand with more depth which methodology has a better performance, two different hierarchical clustering algorithms have been chosen: HDBSCAN and Agglomerative clustering. Both algorithms have been applied using the Dynamic Time Warping (DTW) and Edit distance with Real Penalty (ERP) distance measures. The results have been quantified using the Silhouette coefficient as cluster evaluation technique, and verified using maps with a graphical representations of the clusters. As a test bench for the comparative, a public dataset containing the GPS traces of the trips registered by the users of the transport system of the city of Montreal, Canada; has been used. These trips are labelled by the purpose of the trip, enabling segmented analysis of users movements.

## 2 Related work

Trajectory clustering is a complex problem that has various practical applications such as surveillance security, abnormal behaviour detection, crowd behaviour analysis, or traffic control among others [4].

This work focuses on clustering GPS traces of public transport users. To perform this analysis, it is primarily recommended to carry out a statistical analysis of the data set, such as is proposed in [5] to obtain a detailed description of the data set. Additional statistics can also be obtained, such as common stop sub-sequences or probable destinations from an origin stop.

One of the biggest drawbacks that arise when grouping trajectories is the characteristic nature of the data, since they are composed of GPS locations and timestamps rather than independent variables on an abstract euclidean space. Two potential solutions have been proposed in the majority of cases found in literature. The first involves extracting variables or features from each trajectory, such as length, starting and ending point, etc. In Aaron et al. [6] spatial and usage habit features are extracted from user trajectories. Additionally, Yunzhe et al. [7] employs users trajectories data to determine different users characteristics such as place of residence or conveyance. The second is to apply clustering techniques using distance measures similar to those used in time series analysis, as in Li He et al.[8].Two reviews can be highlighted in the literature regarding trajectory clustering. The first one is about clustering of spatio-temporal data,

which includes trajectory clustering [9]. The second review analyses the distance metrics in the literature used to perform trajectory clustering [10]. In this study, two of those said distance measures, Dynamic Time Warping (DTW) and Edit distance with Real Penalty (ERP), are employed.

## 3 Methodological proposal

The aim of this study is to determine which methodology has a better performance, feature or geospatial trajectory clustering. This section introduces the concept of trajectory, as well as the clustering algorithms and distances used, and defines the metrics employed to determine the result of the comparison.

For feature clustering, the features selected are: origin and destination coordinates, number of points on the trajectory and the distance travelled. For trajectory clustering, the distance measures between trajectories DTW and ERP are tested. Finally, numerical quality results of the clusters are analysed using the silhouette coefficient as a metric. The best results are plotted on a map to verify their coherence.

### 3.1 Trajectory definition

A trajectory is defined as a sequence of geolocated points and corresponding timestamps, which are arranged in chronological order to indicate the subject's movement [3]. The focus of this study is on routes, where each GPS coordinate represents a point on the trajectory. In equation 1, it can be observed how a trajectory is constructed, where $P$ is formed by latitude, longitude, and a timestamp.

$$T = (P_0, P_1, ..., P_i) \tag{1}$$

### 3.2 Clustering algorithms

HDBSCAN and Agglomerative Clustering are be used as clustering algorithms. The reason for using these two algorithms is that they allow the use of different distance metrics than the conventional ones (Euclidean, Manhattan, etc.). This is important because trajectories cannot be measured with these types of distances, as they are a sequence of spatial locations. Therefore, DTW and ERP distance measures are used.

HDBSCAN is an extension of DBSCAN that implements a hierarchical algorithm to establish the maximum distance between neighbours, taking into account the stability of the clusters [12]. DBSCAN is a density-based clustering algorithm that uses the maximum distance between neighbours to determine which instances belong to the same cluster and which ones do not. This algorithm does not assign a cluster to all instances, but some may be classified as noise if they do not have neighbours.

Agglomerative Clustering is a hierarchical clustering algorithm [13], which means that it constructs a tree that has all instances in a single cluster at the root and each instance in a different cluster at the leaves. In this study, an ascending construction has been used, so it starts from the leaves and joins instances until the desired number of clusters is obtained.

### 3.3 Distance measures

Specific distance measures are needed for trajectory clustering. This is because distances like Euclidean or Manhattan measure distances between independent variables. However, when working with trajectories, there is a sequence of stops that can also have a variable length. To calculate these distances, measures that compare these sequences to each other are used, such as DTW and ERP. For this study, an adapted version for trajectories which measures distances in two dimensions instead of one [10] is used.

DTW is a distance measure that, originally, allows comparing two time series by looking for the optimal alignment between them [14]. From this alignment, it measures the distances that exist between them.

ERP is based on edit distance on real sequence (EDR) [15]. This measure calculates the distance between two time series by calculating the number of modifications that would have to be made so that the two signals are equal with a certain tolerance. ERP, in turn, calculates the actual distance necessary to equalize the two series. In case the series do not have the same distance, a reference point is used. It is important to establish an adequate reference point. For this study, the Montreal city geographical centre is used as a reference pint.

### 3.4 Results evaluation

The silhouette coefficient [16] has been selected as evaluation metric. It allows to use both DTW and ERP to calculate the quality of the clusters. In equation 2,it can be observed how the silhouette coefficient of an instance ($i$) is calculated. In the formula, $a$ represents the average distance with the other instances in this cluster, while $b$ represents the minimum distance of the instance with another cluster, which is the closest cluster to the instance being evaluated. To obtain the silhouette coefficient of the clustering, the average of the silhouette coefficient of all instances must be obtained.

$$s(i) = \frac{b - a}{max(a, b)} \tag{2}$$

The representation of the clusters on the maps is drawn using the most central trajectory of the cluster. This trajectory is the one that has the lowest average distance with the rest of the trajectories of the cluster. This is done due to the impossibility of calculating the average of a trajectory and the need to obtain a trajectory that can represent the cluster.

# 4 Dataset description and preprocessing

## 4.1 Description

This study uses the open dataset from the city of Montreal [11] to perform the comparison of algorithms of the two methodologies. This dataset was obtained through a mobile application that records users' trips and inquires about the purpose of the trip at the end of the route. For privacy reasons, the start, and end of the route have been removed in the dataset. Instead, the route begins and ends at the nearest intersection to the start and end of the route. Additionally, this dataset includes only filtered routes, and inconsistent routes have been removed.

## 4.2 Preprocessing

The Montreal dataset consists of 185,285 trajectories, out of which 12,935 belong to public transportation and has been selected for the study. Next, a filtering of the data has been carried out, discarding the trajectories with no defined purpose, which are only 5. To optimize calculations, the number of stops per trajectory have been reduced by only keeping those points that are the closest to a public transportation stop. For this purpose, the distance to the nearest stop of each point is calculated, and then points that are more than 20 meters away from a stop are discarded. Also, if there are multiple consecutive points close to the same stop, the farthest ones are discarded. After this, trajectories that do not pass through at least two stops are discarded, resulting in 6,567 trajectories.

The dataset has been divided into subsets according to their purpose to make the experiment more robust. This results in different datasets that belong to different population segments with different characteristics. Table 1 shows a statistical description of each of these subsets, including the number of trajectories in each subset, the average number of stops, and the average distance travelled.

# 5 Experiments and results

In this section, the results obtained in the experiments are presented and analysed. Two exploring grids of experiments have been designed. The first one, which is used for feature clustering, is a combination of each of the clustering algorithms with each of the subsets. The second one, which is used for trajectory clustering, is a combination of all clustering algorithms with all distance metrics and all subsets of data. Next, the results are compared, and finally, some of the best results are represented on a map for visual inspection.

To understand the results, it is important to take into account the algorithm configuration, specifically the number of clusters used. The configuration used for Agglomerative Clustering is 5 clusters. On the other hand, for HDBSCAN, this number can vary, although the minimum number of instances per cluster is set

Table 1: Number of trajectories, average number of stops of the trajectories and average distance travelled by the trajectories for each of the purposes.

| Purpose | Num. Trajectories | Avg. Stops | Avg. Distance (m) |
|---|---|---|---|
| Back home | 2646 | 8.18 | 6382 |
| Work | 2101 | 8.16 | 6429 |
| Leisure | 602 | 7.83 | 5333 |
| Education | 469 | 7.9 | 7268 |
| Shopping | 323 | 8.17 | 4220 |
| Gastronomic | 160 | 6.83 | 5738 |
| Other | 97 | 8.27 | 6622 |
| Health | 96 | 8.28 | 5689 |
| Picking up a person | 73 | 8.03 | 5327 |

Table 2: Number of clusters generated by HDBSCAN for feature clustering, ERP trajectory clustering and DTW trajectory clustering.

| Purpose | Features | Traj.+ERP | Traj.+DTW |
|---|---|---|---|
| Back home | 2 | 23 | 43 |
| Work | 3 | 23 | 3 |
| Leisure | 2 | 18 | 4 |
| Education | 5 | 17 | 3 |
| Shopping | 3 | 14 | 2 |
| Gastronomic | 2 | 10 | 2 |
| Other | 2 | 9 | 0 |
| Health | 0 | 9 | 0 |
| Picking up a person | 0 | 6 | 0 |

to 5 by default. In order to understand the results of HDBSCAN, Table 2 shows the number of clusters created for feature clustering and trajectory clustering.

After clustering the features using both algorithms and obtaining the Silhouette Coefficients as shown in Table 3, the results are evident. The results achieved through Agglomerative Clustering are stable but relatively low. On the other hand, HDBSCAN exhibits peaks where it produces significantly better results, while in other cases, it fails to create any clusters at all. It is apparent that forming clusters for smaller datasets is considerably more challenging, and if clusters are formed, they tend to be of inferior quality compared to those obtained through Agglomerative Clustering.

Table 3: Silhouette coefficient obtained in the clustering of features

| Purpose | Agglomerative | HDBSCAN |
|---|---|---|
| Back home | 0.16 | **0.40** |
| Work | 0.21 | **0.33** |
| Leisure | 0.20 | **0.23** |
| Education | **0.20** | -0.19 |
| Shopping | **0.18** | -0.10 |
| Gastronomic | **0.22** | -0.01 |
| Other | **0.26** | -0.11 |
| Health | **0.21** | - |
| Picking up a person | **0.24** | - |

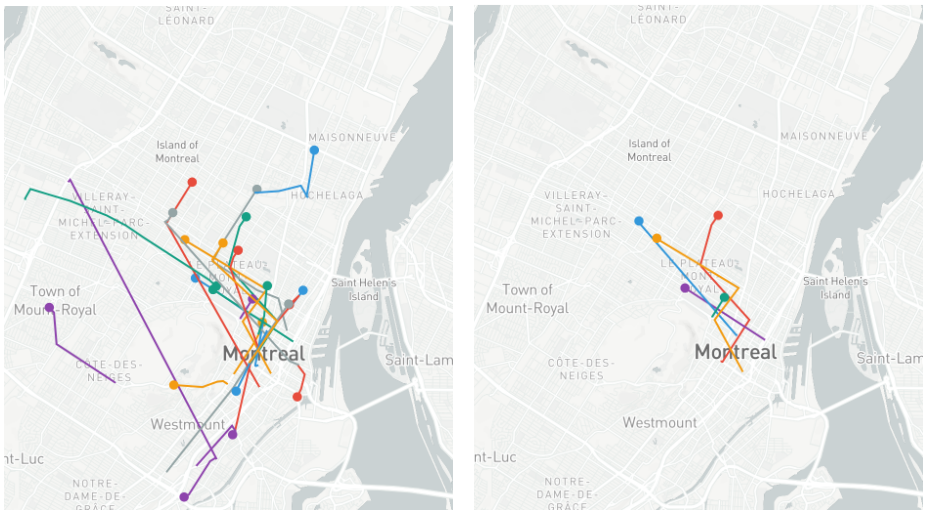
The Silhouette Coefficient results obtained from geospatial trajectory clustering are presented in Table 4. It can be observed that the optimal algorithm combination, in the case of this dataset, is HDBSCAN with ERP. Conversely, the worst performing algorithm combination is HDBSCAN with DTW, which exhibits notably poor results and, in some instances, can not form clusters. Additionally, Agglomerative Clustering performs well with ERP. It is noteworthy that HDBSCAN is optimized for each dataset individually, while Agglomerative Clustering requires optimization of the number of clusters for each dataset. In this study, Agglomerative Clustering uses the same configuration for all datasets. Nonetheless, the results are consistently stable for all purposes. Based on the results obtained from both methodology, it can be concluded that better clusters are obtained in the experiments with geospatial trajectory clustering. Particularly, the results obtained in geospatial trajectory clustering with ERP are significantly better than those obtained in feature clustering. To verify that the results are coherent, some of the experiments that yielded better results are visually analysed.

The maps represent the central trajectory of each obtained cluster; that is, the trajectory with the shortest distance to the rest of the trajectories in the cluster. Each is represented by a circle that indicates the start of the trajectory

Table 4: Silhouette coefficient obtained in the clustering of geospatial trajectories

| Purpose | Agglo. ERP | Agglo. DTW | HDBSCAN ERP | HDBSCAN DTW |
|---|---|---|---|---|
| Back home | 0.61 | 0.34 | **0.99** | -0.46 |
| Work | 0.61 | 0.15 | **0.99** | 0.43 |
| Leisure | 0.62 | 0.28 | **0.97** | -0.26 |
| Education | 0.62 | 0.25 | **0.97** | -0.25 |
| Shopping | 0.62 | 0.31 | **0.92** | -0.04 |
| Gastronomic | 0.63 | 0.18 | **0.84** | 0.08 |
| Other | 0.63 | 0.10 | **0.78** | - |
| Health | 0.63 | 0.38 | **0.78** | - |
| Picking up a person | **0.63** | 0.33 | **0.63** | - |

and a line marking the completed path. In Figure 1, the results of HDBSCAN trajectory clustering and Agglomerative Clustering for the purpose of 'going to work' are shown. It can be observed that HDBSCAN has many more clusters, and they cover a greater variety of trips, while Agglomerative Clustering has fewer and closer clusters. This explains the better results of HDBSCAN due to its greater adaptability.

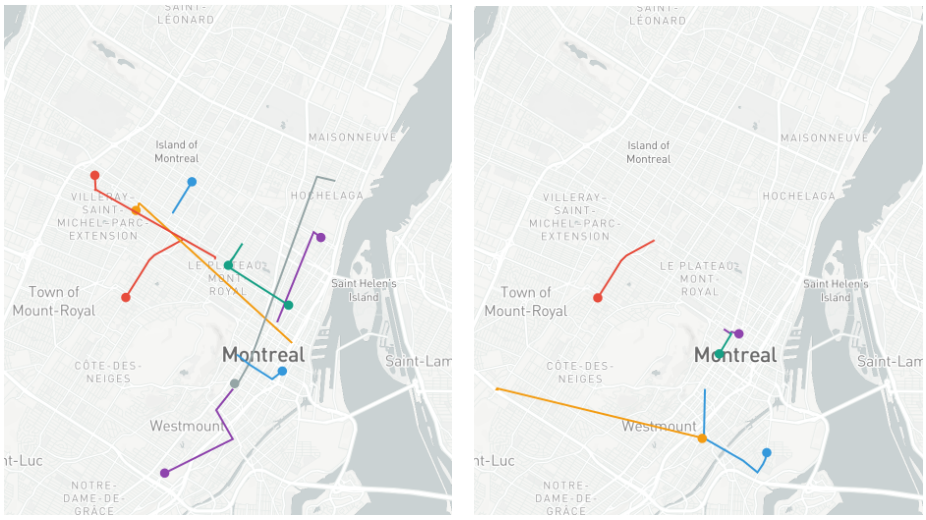



(a) HDBSCAN and ERP (b) Agglomerative Clustering and ERP

Fig. 1: Representation of the trajectories for clusters that aim to go to work

When analysing one of the intermediate values in Table 4, such as gastronomic trips (Figure 2), it is noteworthy that agglomerative clustering results

in more dispersed clusters. However, HDBSCAN has achieved a better representation of the more dispersed clusters. Lastly, when comparing agglomerative clustering with HDBSCAN, agglomerative clustering does not account for clusters that move from the northwest to the southeast, which could be a heavily trafficked area.



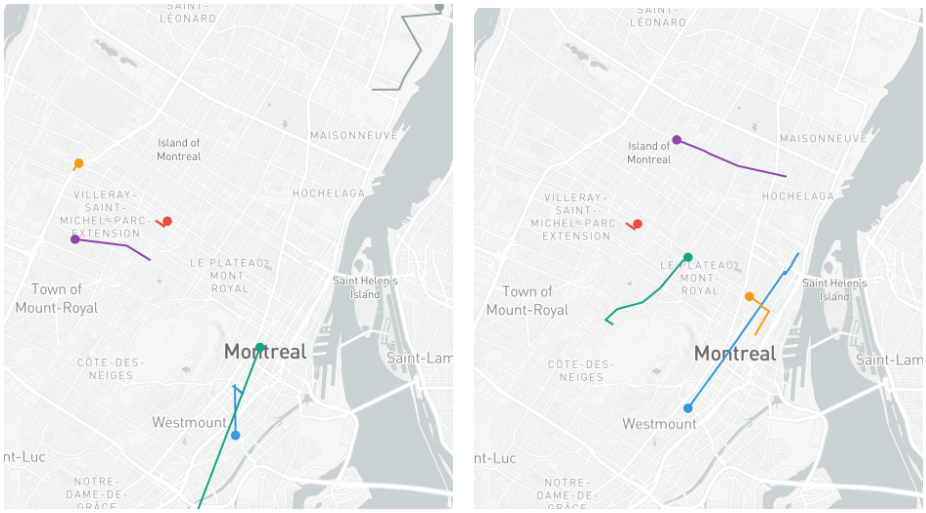

(a) HDBSCAN and ERP (b) Agglomerative Clustering and ERP

Fig. 2: Representation of trajectories for clusters of people moving for gastronomic purposes

In Figure 3, a comparison of HDBSCAN and Agglomerative Clustering is shown, which have very similar silhouette coefficient results. It can be observed that in this case, the difference in the number of clusters is only 1. On the other hand, HDBSCAN results are more dispersed, but in general, the results are very similar in both cases.

## 6 Conclusion and future work

This work presents a comparative for better understanding the clustering algorithms used for clustering routes. In this case, we center on trajectory on public transport in urban environments. The objective is to determine which methodologies and algorithms are more suitable for clustering trajectories. It can be concluded that trajectory clustering offers more solid results than feature clustering. However, it should be noted that this may vary depending on the selected features. On the other hand, the distance measure that offers better results for this type of trajectory is ERP. Lastly, the best clustering algorithm is HDBSCAN, although in some cases it is more unstable, and we cannot determine the




(a) HDBSCAN and ERP (b) Agglomerative Clustering and ERP

Fig. 3: Representation of the trajectories for clusters whose purpose is to go and pick up a person

number of clusters. Meanwhile, Agglomerative Clustering has poorer results, but it has greater stability.

This study presents two principal limitations. The number of clustering algorithms used, which in this case is limited by the need of use of certain distance measures. Also, the study includes a dataset obtained on just one city, as the availability of this kind of datasets is lower than with other problems.

Based on these results, several lines of future work can be studied. On the one hand, the algorithm configurations can be studied deeply, emphasizing the parameter optimization, especially in Agglomerative Clustering. On the other hand, it remains pending of study the applicability of these clusters in tasks such as demand or route prediction. Additionally, a detailed analysis of the characteristics of the clusters could be performed in order to explain and understand what defines them.

## References

1. Beirã, G. & Sarsfield Cabral, J. Understanding attitudes towards public transport and private car: A qualitative study. *Transport Policy*. **14**, 478-489 (2007,11) https://doi.org/10.1016/j.tranpol.2007.04.009
2. Tsafarakis, S., Gkorezis, P., Nalmpantis, D., Genitsaris, E., Andronikidis, A. & Altsitsiadis, E. Investigating the preferences of individuals on public transport innovations using the Maximum Difference Scaling method. *European Transport Research Review*. **11**, 1-12 (2019,12), https://etrr.springeropen.com/articles/10.1186/s12544-018-0340-6

3. Ansari, M., Ahmad, A., Khan, S., Bhushan, G. & Mainuddin Spatiotemporal clustering: a review. *Artificial Intelligence Review*. **53**, 2381-2423 (2020,4), https://doi.org/10.1007/S10462-019-09736-1
4. Bian, J. and Tian, D. and Tang, Y. and Tao, D. A survey on trajectory clustering analysis*arXiv preprint arXiv:1802.06971*. **11**, (2018), https://doi.org/10.48550/arXiv.1802.06971
5. Cogollos, H., Porras, S., Baruque, B., Raffaetà, A. & Zanatta, F. Discovery of tourists' movement patterns in Venice from public transport data. *SAC '22: Proceedings Of The 37th ACM/SIGAPP Symposium On Applied Computing*. pp. 564-568 (2022,4), https://doi.org/10.1145/3477314.3507355
6. Gutiérrez, A., Domènech, A., Zaragozí, B. & Miravet, D. Profiling tourists' use of public transport through smart travel card data. *Journal Of Transport Geography*. **88**, 102820 (2020), https://doi.org/10.1016/j.jtrangeo.2020.102820
7. Liu, Y. & Cheng, T. Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science*. **16**, 76-103 (2020,12), https://doi.org/10.1080/23249935.2018.1493549
8. He, L., Trépanier, M. & Agard, B. Space–time classification of public transit smart card users' activity locations from smart card data. *Public Transport*. **13**, 579-595 (2021), https://doi.org/10.1007/s12469-021-00274-0
9. Ansari, M., Ahmad, A., Khan, S., Bhushan, G. & Mainuddin Spatiotemporal clustering: a review. *Artificial Intelligence Review*. **53**, 2381-2423 (2020,4), https://doi.org/10.1007/s10462-019-09736-1
10. Besse, P., Guillouet, B., Loubes, J. & François, R. Review and Perspective for Distance Based Trajectory Clustering. (2015,8), https://doi.org/10.48550/arXiv.1508.04904
11. Datopian Déplacements MTL Trajet - Dataset. (2017), https://donnees.montreal.ca/ville-de-montreal/mtl-trajet
12. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *The Journal Of Open Source Software*. **2** (2017,3), https://doi.org/10.21105%252Fjoss.00205
13. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *ArXiv E-prints*. (2011,9), https://doi.org/10.48550/arXiv.1109.2378
14. Berndt, D. & Clifford, J. Using dynamic time warping to find patterns in time series. *KDD Workshop*. **10**, 359-370 (1994), https://doi.org/10.5555/3000850.3000887
15. Chen, L. & Ng, R. On the marriage of lp-norms and edit distance. *Proceedings Of The Thirtieth International Conference On Very Large Data Bases-Volume 30*. pp. 792-803 (2004), https://doi.org/10.5555/1316689.1316758
16. Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal Of Computational And Applied Mathematics*. **20**, 53-65 (1987,11), https://doi.org/10.1016/0377-0427(87)90125-7