EARTH OBSERVATION AND GEOMATICS ENGINEERING
website: https://eoge.ut.ac.ir

A dimension reduced clustering approach for the evaluation of trajectory similarities

Amin Hosseinpour Milaghardani[1], Christophe Claramunt[2], Alireza Chehreghan[3]*

[1]Faculty of Civil Engineering and Geodesy, Surveying Engineering Department, Graduate University of Advanced Technology, Kerman, Iran
[2]Naval Academy Research Institute Lanveoc-Poulmic, BP 600, 29240 Brest Naval, France
[3]Mining Engineering Faculty, Sahand University of Technology, Tabriz, Iran

Article history:
Received: 1 March 2020, Received in revised form: 2 February 2021, Accepted: 5 February 2021

ABSTRACT

Nowadays, the very large volumes of trajectory datasets generated by many users and applications offer many opportunities for deriving trends and patterns. Extracting patterns and outliers from people's movements in urban networks is one of the directions worth being explored. For instance, detecting spatial and temporal similarities between trajectory data at different scales and levels of granularity is an important issue. The research developed in this paper introduces a framework based on PCA and K-means methods, and whose objective is to extract similar trajectories from raw trajectory datasets. The approach is first based on a prior characterization of a trajectory with a series of geometric and semantic descriptors. Next, an application of several measures of entropy favors the statistical evaluation of the internal distribution of the main trajectory primitives. Last, and this is the main contribution of this paper, a PCA method is applied to reduce the dimension of the generated primitive data, and finally a K-means clustering technique is used for deriving similarity measures between different trajectories. The whole framework is experimented on top of the Geolife public domain dataset that includes several hundreds of human trajectories in the city of Beijing. The results that emerge show that the whole approach allows for the detection of trajectory similarity patterns using either physical or geometric criteria. Also, similarity detection could be applied for various direction and scales.

KEYWORDS

Trajectory similarity
Spatio-temporal entropy
Geometric and physical descriptor
PCA
K-means.

1. Introduction

Nowadays, trajectory data often available with emerging sensor-based technologies should provide promising opportunities for a better understanding of people's movement in urban environments. While such perspectives open several opportunities for many applications such as traffic management and planning, there is still a need for the development of appropriate data integration, manipulation, and mining techniques. In particular, and in order to develop successful data analysis mechanisms, there is first a need to characterize what is a trajectory and how to model it. The challenge is not so complex at hand and this leads us to informally represent a trajectory by starting and ending points, and the main semantic and geometric primitives that should represent the main internal behavior and dynamic of a trajectory. Another required objective is to reduce as much as possible the data volume embedded by a given trajectory, and this favors further processing steps. Indeed, the objective is to provide a semantic and geometric representation of a given and then a set of trajectories, providing the analytic tools to understand the patterns and trends behind, this possibly providing valuable assets for urban planners and decision-makers (LIN et al., 2014). In fact, such urban patterns, as derived from large trajectory datasets, should reveal urban patterns, the way these trajectories generate trends in space and time, the volumes

* Corresponding author
E-mail addresses: Amin_hoseinpoor@yahoo.com (A. Hosseinpour Milaghardan); Christophe.claramunt@ecole-navale.fr (C. Claramunt); chehreghan@sut.ac.ir (A. Chehreghan)
DOI: 10.22059/eoge.2022.335485.1108

of traffic generated and how those differ in space and time. For instance, detection of high or uncommon traffic patterns will be a valuable contribution for urban infrastructure planning and development as well as for resolving some traffic issues (Aung & Naing, 2014).

The development of appropriate algorithms and data mining processes should consider the spatial, temporal, and semantic dimensions within integrated approaches. The peculiarities that are considered in this paper are similarities and differences that can be exhibited from a large data trajectory repository in order to find similar trajectories according to origin-destination, scale, and direction patterns. This leads us to explore and study the notion of similarity, and this according to the spatial and temporal dimensions considered (Cao et al., 2005). In our previous work, the main intrinsic trajectory characteristics have been identified. Direction, start and end points, sinuosity, curvature, and relative distances have been considered and identified as key spatial primitives (Buchin et al., 2011). Similarly, start and end time, stop points, as well as temporal distances between representative and key internal points of a given trajectory, have been also considered, as well as some derived properties such as velocity and acceleration (Demšar et al., 2015).

Most current methods oriented to the analysis of trajectory similarities can be divided into two groups. A first category applies a systematic evaluation of the distance between the respective characteristic points of two given trajectories (Buchin et al., 2011). A second category decomposes some given trajectories in a series of primitive segments according to some geometric and dynamic parameters and compares them accordingly (Buchin et al., 2012). A trend that also emerges in many studies is the generation of some median trajectories that highlight the main trajectory patterns. A limitation of these approaches is that most of them apply a systematic comparison of the trajectory points, this being expensive from a computational point of view as well as not very efficient as no difference is made between basic primitive points and the ones that embed some noticeable semantic. Moreover, most existing approaches focus on trajectory length, origin, and destination when searching for clusters but not additional trajectory properties. This issue is the main subject of this paper and leads us to introduce a semantic-based approach where first so- critical points of a trajectory are identified according to some geometric and dynamic properties. The distribution of these critical points is quantitatively evaluated by a series of entropy-based measures. In order to reduce the resulting computational complexity of these measures, as well as their legibility, we applied a Principal Component Analysis (PCA) whose objective is to transform the components of a given trajectory towards the most principal components representing the maximum eigenvalues. We postulate that this favors trajectory clustering that will be executed using a K-means method. In abstract, the main goal of the proposed method is to evaluate if the entropy measures of physical and geometric criteria are suitable for similarity measuring of trajectories.

Some advantages of the proposed framework are described as following. First, the proposed method is representing a spatial-temporal distribution-based measure considering physical and geometric descriptors. So, it makes possible the similarity comparison of trajectories from one or some descriptors separately or combinatory. As a result, comparison of too sophisticated trajectories becomes possible without regarding the functions that are used in previous studies. Furthermore, a positive point of proposed method is detection of similar geometric trajectories in scattered direction, start and end points and distances. So, their results can be used in detection of geometric patterns in different scales and directions

The rest of the paper is organized as follows. First section 2 reviews related work while section 3 introduces the main principles of our framework. The approach is evaluated by a report on the implementation developed so far in section 4. Last Section 5 concludes the paper and draws some perspectives for further work.

2. The Related Work

Over the past few years, the analysis and search for trajectory similarities based on physical and geometric descriptors have been the object of several research works. According to (Parent et al., 2013), stop points provide for instance valuable inputs for studying and differencing trajectory data. In fact, a key issue relies on the identification of effective descriptors. Events and activities associated with either stop points or movements can give useful insights for studying trajectory differences and similarities (Asakura & Hato, 2004; Hofmann et al., 2009; Hornsby & Cole, 2007; Lee et al., 2011; Lee et al., 2008; Pelekis et al., 2009; Perttunen et al., 2015; Robinson et al., 2017; Zheng et al., 2010; Zhou et al., 2015). When considering geometric properties, structuring a trajectory by segments based on curvature points has been suggested as a valuable method for identifying the main characteristics and then facilitating the search for trajectory patterns (Bashir et al., 2007; Harguess & Aggarwal, 2009; Himberg et al., 2001; Kafkafi & Elmer, 2005; Kafkafi et al., 2009; Soleymani et al., 2014). Additional parameters such as velocity (Asakura & Hato, 2004; Dodge et al., 2009; Fang et al., 2009; Lu et al., 2015; Soleymani et al., 2014; Zheng et al., 2010), direction (Asakura & Hato, 2004; Aung & Naing, 2014; Gao et al., 2013; Lee et al., 2008; Lu et al., 2015; Monreale et al., 2009; Pelekis et al., 2009; Perttunen et al., 2015; Zheng et al., 2010), turning points and angle (Dodge et al., 2009; LIN et al., 2014; Monreale et al., 2009; Soleymani et al., 2014), acceleration (Dodge et al., 2009; Dodge et al., 2011; Zheng et al., 2010), sinuosity (Aung &

Naing, 2014; Dodge et al., 2011; LIN et al., 2014; Soleymani et al., 2014), distance (Asakura & Hato, 2004; Cao et al., 2005; Dodge et al., 2009; El Mahrsi & Rossi, 2012; Fang et al., 2009; Gonzalez et al., 2008; Lee et al., 2011; LIN et al., 2014; Morzy, 2007; Pelekis et al., 2009), travel time surely provide additional insights (Dodge et al., 2008; Giannotti & Pedreschi, 2008). When considering large trajectory datasets, searching for outliers that deviate from median trajectories in both space and time has been studied in related work (Dodge et al., 2009; Laube & Purves, 2011). One common difficulty of all these methods appears at the computational level, especially when dealing with large trajectory datasets.

In order to improve processing times, a given trajectory should be filtered by keeping the most relevant points according to the most relevant geometric descriptors. Several algorithms have been already explored to do so using spatial and temporal descriptors such as turning points, directions, sinuosity, and speed (Buchin et al., 2011; Dodge et al., 2009; Lin & Hsu, 2014). A key issue when applying a filtering algorithm to a given trajectory is the identification of the most relevant geometric descriptors, the ones that make sense with respect to the application domain considered, as well as avoiding dependent parameters. For instance, curvature and direction, speed and acceleration, are dependent descriptors that should not be considered together. Most of the studies we are aware of and mentioned in this section show that almost all of these methods are oriented to the detection and extraction of movement pattern process all trajectory points or applied some geometrical filters that do not take into account the whole semantics of the considered trajectories. This leads us to propose an approach that first decrease the number of trajectory points during the analysis process to a meaningful level of critical points. The objective is to decrease processing time by reducing trajectory data volumes. The second peculiarity of our method is that it combines a series of geometric and physical descriptors that also considers both the spatial and temporal dimensions.

3. The Proposed Method

Overall, two important issues are hereafter considered as fundamental assumptions for the development of our approach. The first one is the identification of the minimum and most relevant geometric descriptors to characterize a given trajectory. The second one is to apply a statistical evaluation of the internal distribution of the main trajectory primitives using spatio-temporal entropy measures introduced in our previous work (Hosseinpoor Milaghardan et al., 2018b). Next, we introduce a PCA method to reduce the dimension of the generated primitive data as the number of included physical and geometric criteria will increased. Moreover, there is a significant demand to detect and find the most affecting criteria on separating the trajectories. and finally, a K-means clustering technique for deriving similarity measures between different trajectories. It is an unsupervised clustering method that can effectively find the most number and distribution of trajectory divisions. Figure 1 summarizes the different components of the whole methodological approach.

3.1. Critical Points Detection

The first objective is to derive a minimum number of critical points while considering the physical and geometric characteristics of trajectory data. Geometric descriptors include curvature, turning, and self-intersection points. The first two parameters show the shape and geometry of trajectory while the self-intersection parameter is applied for detection and separation of self-intersecting points. Physical descriptors include stop status or user movement and velocity. We apply a convex hull structure on trajectories that for any curvature on that trajectory a Convex Hull (CH) structure is formed. In the example presented in figure 2 Convex Hull are formed such as $CH1=\{P_1,P_2,P_3,P_4,P_5,P_6,P_7\}$,$CH2=\{P_7,P_8,P_9,P_{10},P_{11},P_{12}\}$, $CH3=\{P_{12},P_{13},P_{14},P_{15},P_{16},P_{17},P_{18}\}$, and $CH4=\{P_{18},P_{19},P_{20},P_{21},P_{22},P_{23},P_{24},P_{25},P_{26},P_{27},P_{28}\}$.

Turning Points: Differences between a straight trajectory and a complex one can be revealed by differences in curves. Keeping the main geometric properties of such trajectories requires a series of parameters that can show the position, number, and form of these curves.

Curvature Points: Each curve identified in a given trajectory encapsulates primitive features and forms that can be studied using the distance between turning points and sinuosity.

Self-intersected trajectories: Self-intersections generally arise in many trajectories, but one should make a difference between noisy self-intersections that should be eliminated, and larger self-intersections that are common in many contexts such as maritime lines or animal migrations to mention some basic examples. This is why appropriate filtering mechanisms should be designed to detect them (Hosseinpoor Milaghardan et al., 2018a).

Stop points: When considering time as an additional dimension stop points are mandatory primitives to take into account as these often denote some specific activities. Detection of these stop points is based on an application of Dempster-Shafer's theory and belief and non-belief functions (Milaghardan et al., 2018).

Speed and acceleration: Velocity is derived for all relevant trajectory points as well as accelerations, these providing additional parameters of interest to take into account.

All the above properties are detected and saved accordingly as additional parameters of a given trajectory, and this is at the trajectory point levels. A comprehensive

graph is derived from the starting to the ending points and along the trajectory respective primitive points associated with the primitive attributes mentioned above. The nodes of this graph include geo-referenced critical points while edges keep spatial and temporal distances between them.

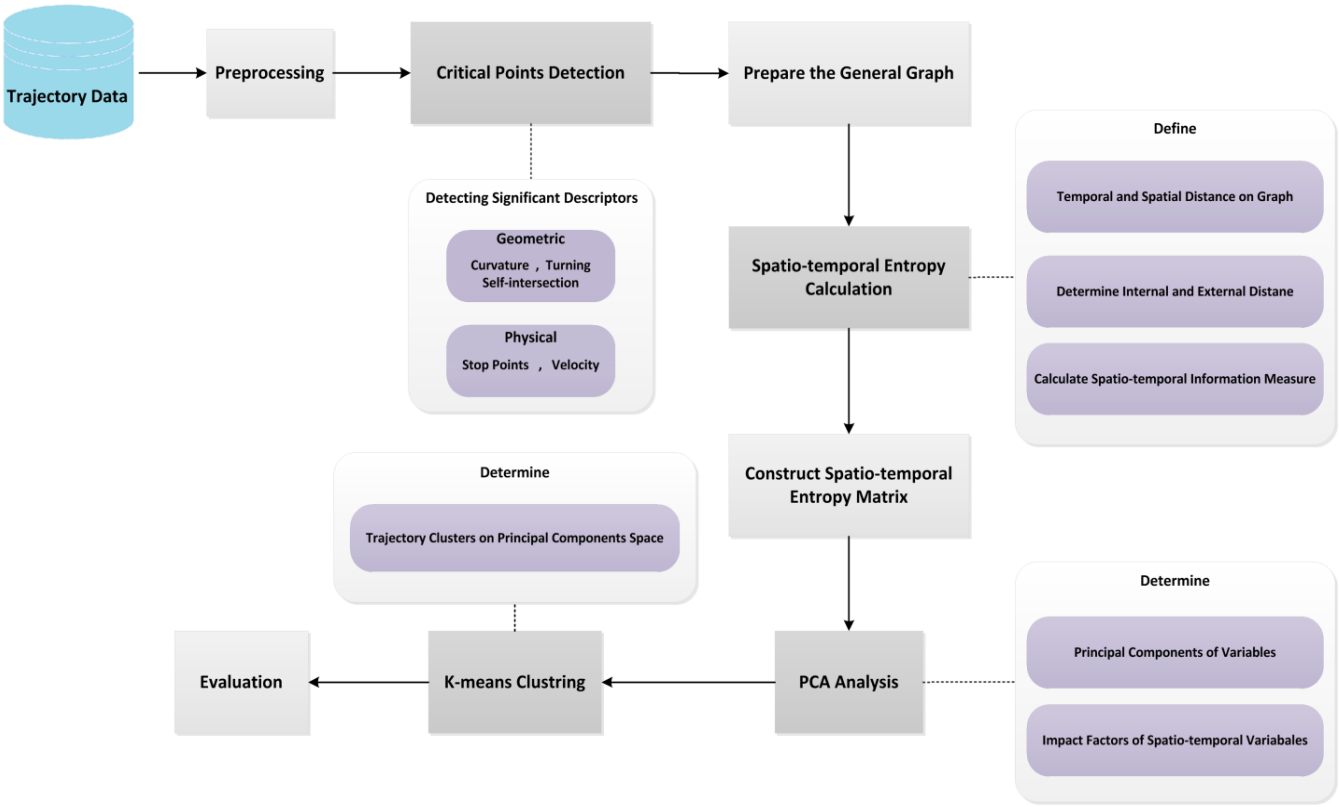


Figure 1. Method flowchart

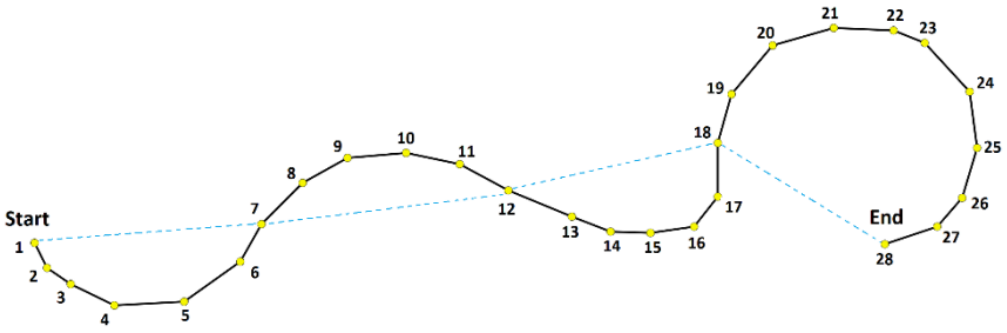


Figure 2. Trajectory convex hull

3.2. Spatial-temporal Entropy calculation

Let us develop the principles of the entropy method and how it can be used to detect trajectory similarities according to the critical points and associated parameters considered as introduced in our previous work (Hosseinpoor Milaghardan et al., 2018b). First the measures of spatial (H_S^i) and temporal (H_T^i) entropies are defined as follows (Hosseinpoor Milaghardan et al., 2018b).

$$H_S^i = -\sum_{i=1}^{n} \frac{d_i^{int}}{d_i^{ext}} p_i \log_2(p_i) \quad (1)$$

$$H_T^i = -\sum_{i=1}^{n} \frac{td_i^{int}}{td_i^{ext}} p_i \log_2(p_i) \quad (2)$$

The inner distance of class i denoted d_i^{int} represents the average of distances between entities of class j. Similarly, the external distance d_i^{ext} and represents the average of distances between entities of class i and entities from other classes.

When applying the spatio-temporal entropy to the cross-analysis of the properties of two given trajectories, every physical and geometrical parameter, previously introduced in section 3.1, is first associated with spatial and temporal entropy measures. Therefore, trajectory clusters are derived according to similar entropy values. The next goal is to introduce a spatial-temporal entropy matrix for all trajectory data. In order to give a relatively complete view of the problem, the temporal and spatial entropy values of each of the trajectories for the different classes are schematically presented in this matrix. The dimension of this matrix is

$(2m + 2) \times n)$ where n denotes the number of trajectories considered, while m denotes the total number of semantic and geometrical parameters, and T_1, T_2, …, T_n represent the id of trajectory 1, trajectory 2, and trajectory n, respectively (Table 1). Let us introduce an example of derived entropy values of entropies for the two trajectories 47 and 56 extracted from the sample dataset (Figure 3).

The results presented in table 2 can be used to describe the semantics aspects of the trajectories 47 and 56, and for further comparison based on the predefined parameters. For example, the results show a close similarity when considering the temporal dimension (i.e., a small difference in temporal entropies 0.335 and 0.359 as well as for curvature values 0.32 and 0.27 trajectories for trajectories 47 and 56, respectively), while not when considering the spatial dimension (i.e., the higher difference in spatial entropies 0.376 and 0.287; curvature values 0.53 and 0.18 for trajectories 47 and 56 respectively).

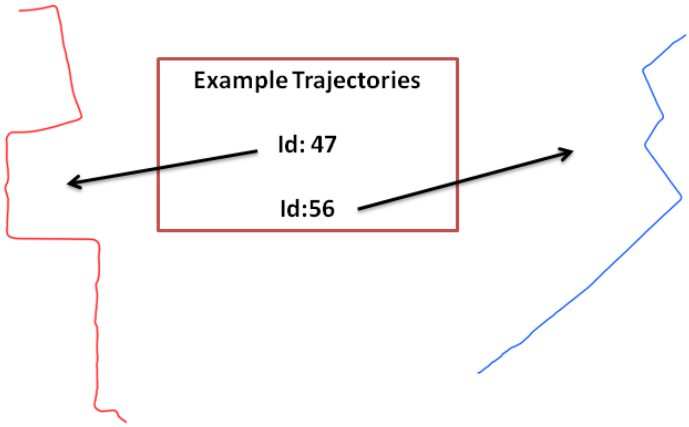


Figure 3. Example trajectories 47 and 56

Table 1. Spatial-Temporal entropy matrix

			T_1	T_2	T_3	. . .	T_n
Spatial	Spatial Entropy		$V_{1,1}$	$V_{1,2}$	$V_{1,3}$	. . .	$V_{1,n}$
	Physical measure	speed	.				.
		stop	.				.
	Geometric measure	Curvature	.				.
		Turning					
		Intersection					
Temporal	Temporal Entropy						
	Physical measure	speed					
		stop					
	Geometric measure	Curvature					
		Turning					
		Intersection	$V_{12,1}$	$V_{12,2}$	$V_{12,3}$	. . .	$V_{12,n}$

Table 2. Matrix example for trajectories 47 and 56

			Trajectories	
			47	56
Spatial	Spatial Entropy		0.376	0.287
	Physical measure	speed	0.34	0.51
		stop	0.11	0.14
	Geometric measure	Curvature	0.53	0.18
		Turning	0.38	0.46
		Intersection	0	0
Temporal	Temporal Entropy		0.335	0.359
	Physical measure	speed	0.29	0.53
		stop	0.36	0.28
	Geometric measure	Curvature	0.32	0.27
		Turning	0.41	0.49
		Intersection	0	0

3.3. Dimension Reduction

Due to the large number of features embedded in a spatial-temporal entropy matrix, ineffective features should be identified in order to reduce the features space dimension. We applied a PCA in which each of the considered parameters, along with their critical points, is considered as a vector. The purpose is to find alignments in space in which these parameter vectors have the most relative variance. For this purpose, the eigenvalues of this matrix are extracted from a covariance matrix and the main components are prioritized according to the highest obtained eigenvalues. The objective is to search for a linear function $a_1'X$ of the elements of X having a maximum variance, where a_1 is a vector of p constants $a_{11}, a_{12}, \dots, a_{1p}$ and ' denotes transpose, so that (Jolliffe, 2011)

$$a_1'X = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = \sum_j^p a_{1j}x_j \qquad (3)$$

where the respective a_1 parameters denote the respective values of the geometric and physical parameters introduced in the previous section.

One important goal is to find the correlation between the parameter vectors for each considered class. A squared score matrix is used to determine the coefficient of influence of each of the parameters in the main components. In fact, one of the goals is to find the number of main components that represent the most variance between the considered classes. The eigenvalues of the principal components can be used as an enhancement parameter to select the most detective components. The components with the largest values of eigenvalues are selected as principal components. The eigenvalue can be described as follows:

$$a_1' \sum a_1 = a_1' \lambda a_1 = \lambda a_1' a_1 = \lambda \qquad (4)$$

where λ is an eigenvalue of $\sum$ and a_1 is the corresponding eigenvector. A component can be considered as a principal component when the related eigenvector is as large as possible (Jolliffe, 2011).

The higher the number of these components, the lower the class dependencies will be. A square cosine coefficient denotes differences between the derived coefficients of the aforementioned classes for each component. The next step is to select the number of sufficient numbers for components. The cut off condition is described as follows:

$$v > \frac{v_{max}}{10} \qquad (5)$$

where v denotes the variability of the considered component and v_{max} the maximum variability. A small cut off value as related to v_{max} is selected as the first component exhibits around 90% variability, so small variability values will not help to discriminate the considered objects.

3.4. Similarity Detection

The final part of this approach is the extraction of similar patterns among trajectory data. Two important issues and initial conditions are considered. First, the number of hidden patterns in the trajectories should be unclear, and the search for patterns must be non-supervised. The proposed STE-SD structure applies a K-means clustering method to detect trajectory similarities. The K-means algorithm takes an input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high, but the inter-cluster similarity is low. Cluster similarity is measured with respect to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or the center of gravity (Miller & Han, 2009). The K-means algorithm proceeds as follows. First, it randomly selects k objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, a square-error criterion is used, and defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2 \qquad (6)$$

where E is the sum of the square-error for all objects in the data set; p is the point representing a given object; and m_i is the mean of a cluster c_i (both p and m_i are multidimensional).

The objective is to achieve the best clustering, in order to reduce the variance of the distance within the clusters and to increase the variance of the distance between the clusters that can be derived using Eq.6. The number of clusters has been specified using the clustering parameters including in-class and between-class variance in order to select the optimum value. The clustering is performed in the reduced dimension space (i.e., that results from the PCA) derived from a spatial-temporal entropy matrix. One of the properties of the proposed method is the ability to extract some patterns based on different parameters. In other words, the K-means method, as derived from the spatial-temporal entropy matrix, can extract some patterns using parts of all of the spatial and temporal properties. Therefore, by identifying the main components, clustering is done separately and with arbitrary variables. Accordingly, it is possible to study and extract some temporal and/or spatial patterns, as well as their relations to each. This allows for the analysis of similarities between different patterns. For example, this feature can be used to extract temporal-geometric patterns or spatial-physical locations. Each of these patterns can be used according to the application and needs of studies with less computational volume and higher accuracy.

4. Experiments

Let us introduce the implementation experimented so far. We first describe the dataset used and then the results of the implementation. The principles of our approach are applied

to a large urban trajectory dataset available in the city of Beijing. The Geolife project collected a large repository of urban trajectories recorded by taxis, buses, or even human beings equipped with GPS receivers from 2007 to 2012 (Zheng et al., 2009). The main advantage of this reference dataset is that is fully available, and it has been largely used as a benchmark database for further research and studies. For the objective of our experiments, we selected a sample of this database made of 326 trajectories that reflect a relatively large variety of human displacements made either by taxis, buses, or even walking as presented in Figure 4. These trajectories overall represent 83412 trajectory points and a total distance of 672195 m. The shortest trajectory covers a distance of 8.54 m while the longest one is 14408.2 m, the mean length of these trajectories is 2417.97 m. Likewise, the mean sampling distance covered between two trajectory points is 10.21 m and the mean sampling time is 5.11 seconds.

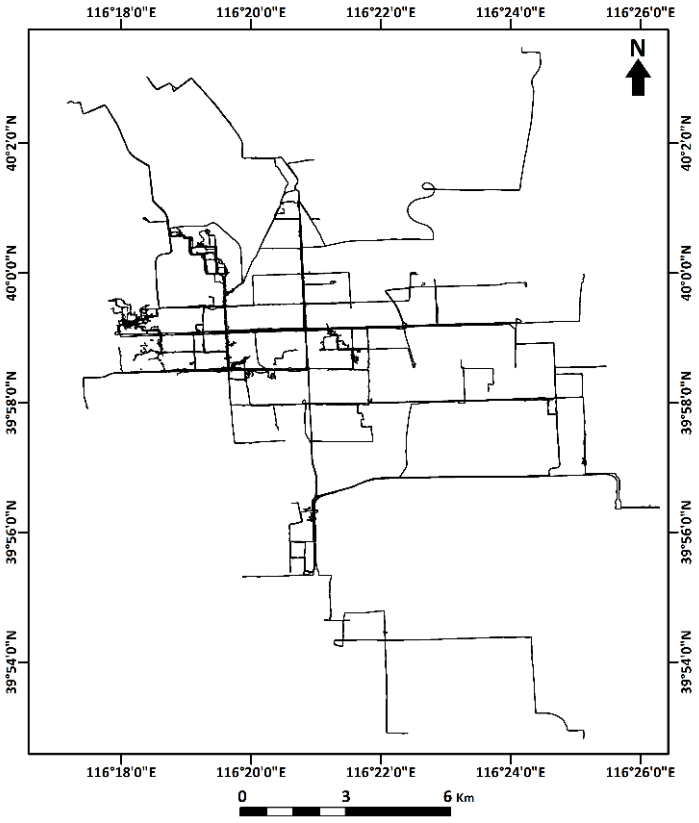


Figure 4. 326 trajectories selected for the implementation

4.1. Implementation

According to the explanations provided in the methodology section, the proposed STE-SD method involves four main steps and is hereafter presented in four separate sections. In the first step, critical points are identified for each physical and geometric descriptor, and their results are used as input data for the second step. Further, taking into account the equations presented in the second section, spatial and temporal entropies and information criteria are calculated and a spatial-temporal entropy matrix is derived. In the third step, the space of the matrix variables is reduced to the core component space. At the final step, similar spatio-temporal patterns are identified. The evaluation of the results is presented at the end of this section.

4.1.1 Critical Points Detection

A convex hull geometric structure has been implemented for 326 trajectories and gave 7498 convex hulls. In fact, the generation of convex hull structures implies specific primitive values that differ from other trajectory properties. For example, trajectories 64 and 68 are geometrically similar but have 76 and 92 convex hull structures, respectively (Figure 5) while the start and end points of these two trajectories are similar. However, convex hulls with either a distance lower than 0.02D (D denoting the trajectory length) and convex hulls with less than four points are removed from the resulting convex hulls. Accordingly, 2317 structures were removed from the generated convex hulls. Table 3 shows the number of deleted convex hulls per trajectory category together with their critical points. The results highlight the number of deleted convex hulls per category of trajectories, with the largest numbers for trajectories with lengths between 800 and 1200 meters.

Next, critical points for speed that denote speed changes

apart from stop points were further considered. This overall gives 5720 points. The distribution of these points according to different trajectory categories is presented in Figure 6.

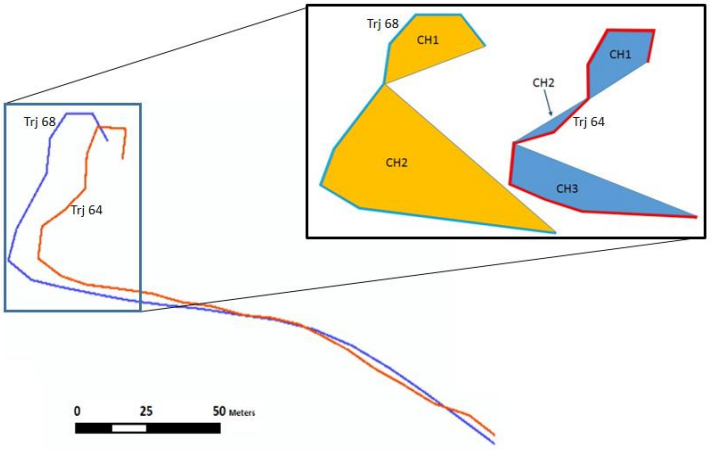


Figure 5. Similar trajectories 64 and 68 but with different convex hulls

Table 3. Filtering of trajectory convex hulls.

Properties	**Category**			
	1	2	3	4
Length of trajectory	0-1000	1000-4000	4000-8000	8000-15000
Number. of Trajectories	56	82	127	61
Number. of Primary CH	510	1245	3910	1833
Number. removed CH	63	235	1376	639
The variance of Distance to CH line	3.41	7.33	14.57	17.20
Number. of Curvature points	447	1010	2534	1194
Number. of Turning points	449	1012	2536	1196
Number. of Intersection points	5	26	58	31

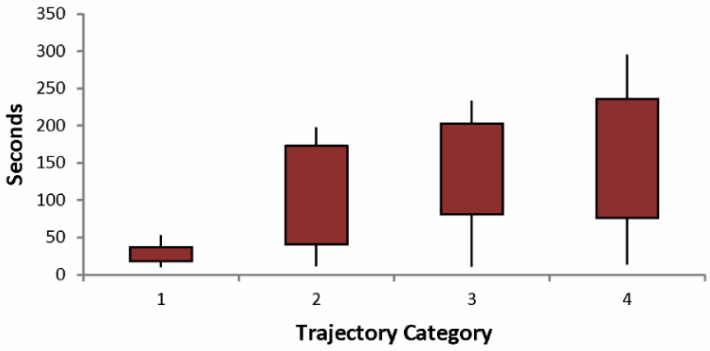


Figure 6. Box Plot graph of the temporal distribution of speed critical points per category

Figure 6 shows that there are substantial differences between speed behaviors over time when considering trajectory lengths. The minimum temporal difference is exhibited in group 1 with values from 18 to 37 seconds while the maximum time interval is obtained in group 4 with values of 76 to 236 seconds. The next physical parameter considered is the status of the movement of the trajectory points, in other words, the identification of the stop points. The detection of stop points is far from being a straightforward task when especially considering the notion of uncertainty. In related work, we introduced an approach based on the Dempster-Shafer theory of evidence, and

whose objective is to detect trajectory stop points and associated degrees of uncertainty (Milaghardan et al., 2018). The minimum, maximum, and average values of the Belief, disbelief, and uncertainty of all trajectory points including the identified stop points are given in Table 4. The maximum values of belief denote a high probability of stop points, while maximum values of disbelief show probable movement points. Note that high uncertainty values denote unknown situations for the considered points.

Table 4. Belief, non-belief, and uncertainty values for identifying stop points

	For all points			For Stop points		
	Minimum Value	Maximum Value	Average	Minimum Value	Maximum Value	Average
Belief	0.09	0.945	0.864	0.723	0.945	0.834
Disbelief	0.02	0.894	0.448	0.12	0.27	0.145
Uncertainty	0.04	0.23	0.135	0.06	0.19	0.125

4.1.2 Spatial and Temporal Entropies

The first step for deriving spatial and temporal entropy values is to obtain the spatial and temporal distances between critical points, as well as the five intended physical and geometrical parameters. Therefore, spatial distances of 326 studied trajectories were calculated as the Euclidian distance between consecutive critical points of each parameter, as well as temporal distances (Figure 7). This figure shows the direct relationship between temporal and spatial values for all parameters except for the stop parameter, which includes the maximum spatial distance of 1073.6 meters and the lowest temporal distance of 68 seconds.

Calculation of the information criterion for each of the parameters requires obtaining their internal and external spatial and temporal distances. Therefore, using the spatial and temporal distances of the critical points of the parameters, as well as the internal and external distances were calculated. The average results of these calculations for the parameters of 326 studied trajectories are presented separately by the time and spatial values in Table 5.

One of the important trends that appears from the spatial values in Table 5 is the difference between the average internal distance, and the average distance of consecutive critical points presented in Tables 6 and 7. The values of Table 4 show the spatial distance between two consecutive critical points of each class, while the internal values of Table 5 show the average distance between any point of the class and other points of that same class. The external spatial values show the average distance between the points of each class and the other points. Table 5 shows that there is a mean of 1099.7 meters between every two considered stop points while the temporal distance mean is 672.0 seconds. This overall shows a 10min average per km this being a relatively slow value for buses and taxis while acceptable for walking patterns. The figures that appear from the critical points show the respective complexity numbers of the different physical and geometrical parameters according to their spatial and temporal dimensions as well as the fact that the two are not completely correlated. After calculating the values of the internal and external distances for the different parameters, the spatial and temporal information criterion of each of them is obtained. Spatial and temporal entropies are calculated independently for each trajectory. After calculating entropy and information criteria for the desired data, the spatial-temporal entropy matrix is formed, which is then used as input for the next step.

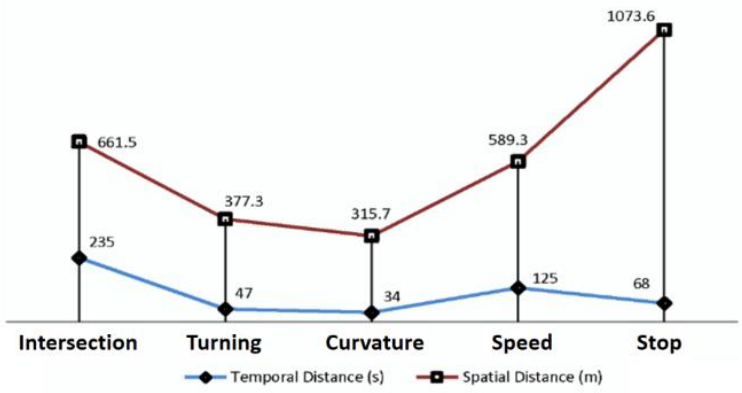


Figure 7. Relationship between spatial and temporal distances of the physical and geometrical descriptors

Table 5. Average of internal and external spatial and temporal distances

		Physical Parameters		Geometric Parameters		
		Stop	Speed	Turning	Curvature	Intersection
Spatial	Internal Distance	1099.7	1425.2	1669.5	1351.8	744.5
	External Distance	1733.9	1886.5	1905.2	1922.8	2185.3
Temporal	Internal Distance	672.0	844.7	1297.1	983.2	406.3
	External Distance	2184.9	4521.6	6621.3	5447.0	8341.6

4.1.3. Dimension Reduction

The third part of the proposed method consists of the dimension reduction of the spatial-temporal features of the spatial-temporal entropy matrix which was described in the previous section. The proposed method applied is the PCA method. Therefore, 96 trajectories of the sample data with various lengths were used to identify the coefficient of variable influence. The chosen variables for the desired matrix include 6 spatial and entropy criteria, 6 temporal and entropy criteria, and the number of trajectories CHs, making overall thirteen variables. Due to the variable distributions, the Pearson coefficient was used for the PCA method implementation. Table 6 shows the results of the calculation of eigenvalues for the obtained components.

The results that appear in Table 6 show that by using the first two components, 85.67%, and by using the first three components, 95.95% of data variability can be shown. The first component has the largest share of 67.29% in data variability. In order to illustrate the cumulative variability, the components of the screen plot graph of eigenvalues and cumulative variability of them are shown in Figure 8.

Given that 95.95% of the variability of the considered data can be identified using the first three components, they are considered main components. Also, in order to evaluate the relationship between these components, their correlation circle diagrams are shown in Figure 9. This Figure includes three diagrams, which show the relationship between the components pairwise.

Table 6. Eigen Values and Variability calculated for main components

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Eigenvalue	9.421	2.570	1.442	0.365	0.112	0.044	0.014	0.013	0.007	0.005	0.004	0.002
Variability (%)	67.295	18.354	10.300	2.607	0.799	0.317	0.102	0.091	0.051	0.037	0.030	0.013
Cumulative %	67.295	85.649	95.950	98.556	99.355	99.672	99.774	99.865	99.916	99.953	99.983	99.996

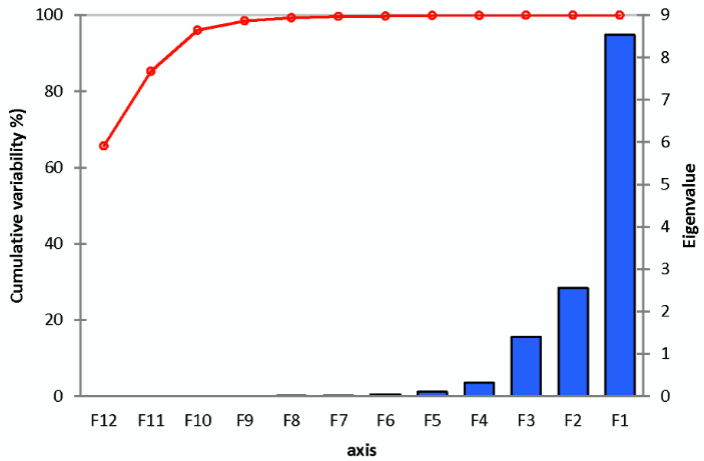
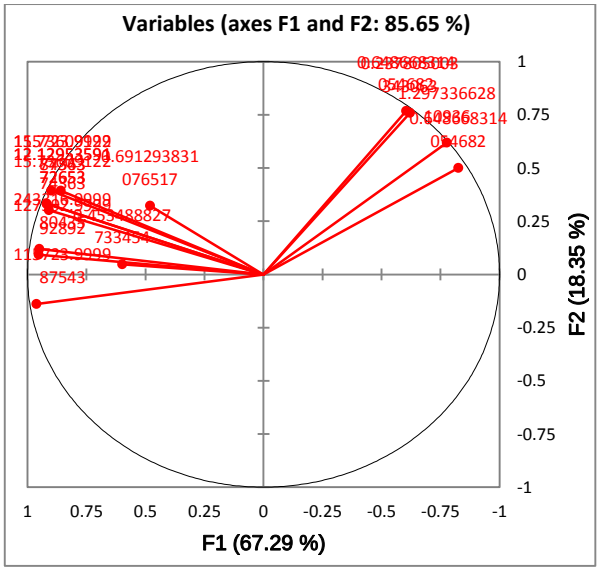
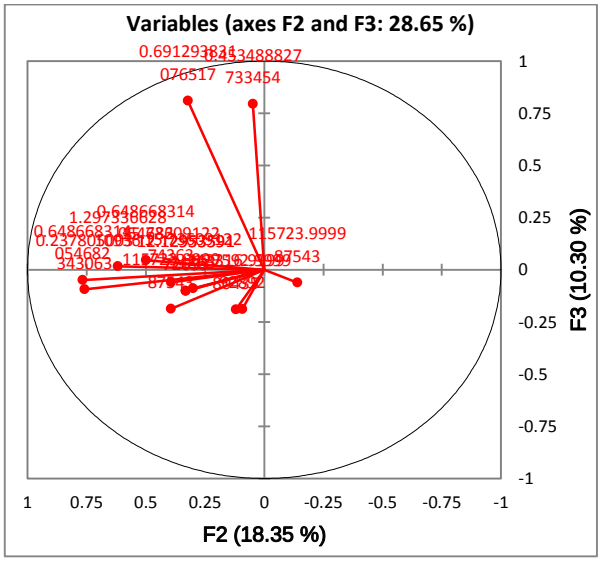


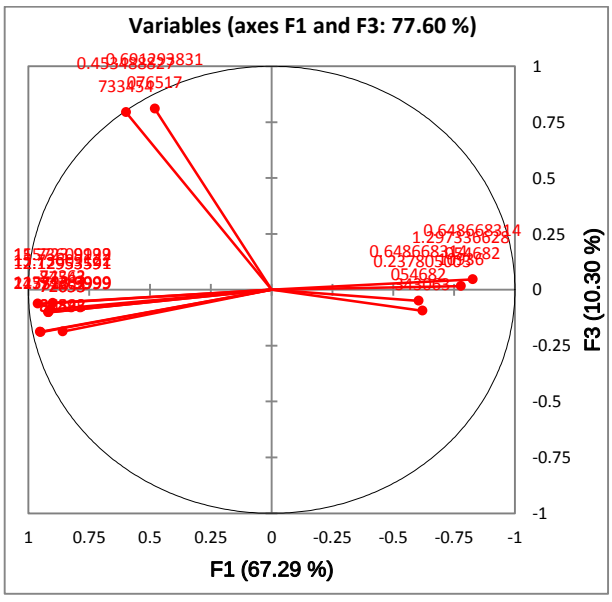
Figure 8. Screen plot graph for showing cumulative variability of the main components



a.



b.



c.

Figure 9. Correlation circle diagram for (a) first and second components; (b) second and third components; (c) first and third components.

As seen in the correlation circle diagrams presented in Figure 9, the highest non-correlation was calculated between F1 and F2 components with a value of 85.65%, and the lowest was calculated between F2 and F3 components with a value of 28.65%. It can be said that the first two components are more capable of displaying variabilities, but in order to increase the precision and more accurate identification of the variable relationships, the third component is also used. In order to more accurately examine the ability of these three components, Figure 10 presents the number and volume of the trajectories and their displayable changes in two axes of F1, F2, and F3, per-wisely. As it is seen in this figure, the best distribution of data is in the direction of the axis of the F1 and F2 components with the dense point distribution of the points with the least connection between the axes F2 and F3.

In order to identify the representation quality of each variable, squared cosines values are calculated for each of them, which are presented in Table 7. In fact, this table shows the weight and coefficient of participation of each of the variables in the obtained components. The values of this table are between zero and one, and the closer the number is to one, the greater the impact of the corresponding variable.

In order to detect valuable variables, the values of squared cosines higher than 0.5 are shown in dark colors. It appears that some variables have values less than 0.5 in Table 7. These variables include the spatial information of the stop criterion and temporal information of speed, curvature, and turning information criteria, respectively. Temporal and spatial intersection parameters have lower than 0.1 values as shown in Table 7, and this provides insufficient impact as principal components. Another parameter that has relatively high values of square cosine in F1 and F2 components is the number of convex hull structures as declaring a meaningful relationship when discriminating the trajectories geometry. As a whole, this reveals that spatial parameters have a greater effect than the time parameters when categorizing trajectory data since the considered Geolife data set includes different modes of movement. It appears that the temporal parameter could be used for trajectory categorization while using a unique mode of transportation.

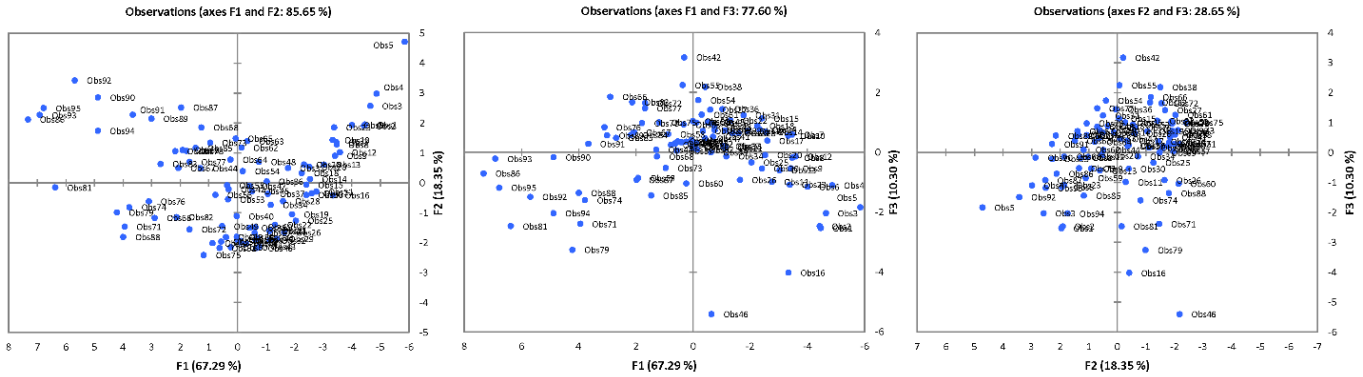


Figure 10. Distribution of trajectories in the first three components

Table 7. Square cosines coefficients calculated for main components

Variable		F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Spatial	Speed	0.604	0.383	0.000	0.000	0.007	0.002	0.000	0.001	0.000	0.000	0.001	0.000
	Stop	0.365	0.589	0.002	0.017	0.022	0.000	0.001	0.002	0.000	0.000	0.001	0.000
	Curvature	0.682	0.251	0.002	0.007	0.057	0.000	0.000	0.000	0.000	0.000	0.001	0.000
	Turning	0.810	0.155	0.004	0.026	0.001	0.000	0.001	0.001	0.000	0.001	0.000	0.001
	Intersection	0.044	0.110	0.010	0.034	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
	Entropy	0.827	0.091	0.008	0.068	0.000	0.001	0.001	0.000	0.002	0.001	0.000	0.000
Temporal	Speed	0.844	0.110	0.010	0.034	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
	Stop	0.231	0.104	0.655	0.003	0.000	0.000	0.004	0.002	0.000	0.000	0.000	0.000
	Curvature	0.359	0.002	0.630	0.002	0.000	0.000	0.004	0.002	0.000	0.000	0.000	0.000
	Turning	0.385	0.577	0.009	0.014	0.010	0.000	0.002	0.003	0.000	0.000	0.000	0.000
	intersection	0.026	0.019	0.004	0.031	0.000	0.018	0.001	0.000	0.000	0.001	0.000	0.000
	Entropy	0.607	0.009	0.036	0.042	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000
Convex Hulls		0.738	0.255	0.035	0.045	0.006	0.020	0.000	0.000	0.001	0.000	0.000	0.000

4.1.4. Similarity Detection

Similarities between data in the dimension reduced space are explored using the PCA method. By applying the K-Means technique, each trajectory is depicted in the space of the previously mentioned three main components with the values obtained from the PCA method, and then similar trajectories are identified using the K-Means technique. In Figure 11 all intended 95 trajectories are shown in the 3-D space of the three main components extracted from the PCA method. Figure 11 shows a 3D presentation of the three diagrams presented in figure 10 and where the maximum variability is derived for components F1 and F2.

The K Means method is implemented on the values depicted on the three main components' space, considering 4 to 8 clusters. The determinant of the within-class variance is used as the main parameter for clustering. The results of within class variance for each number of classes are shown in Figure 12 as we are searching for the class number with minimum within-class and maximum between-class variances.

Similarly, the obtained results of within-class and inter-class variances are shown in Table 8. Among these, 8 classes are grouped according to less within-class variance and higher inter-class variance. In other words, these 8 clusters group trajectories with minimum differences regarding their respective spatio-temporal entropy values.

The distances between cluster centers in 8 class cases have the highest obtained values. Table 9 presents the matrix of distances, and also trajectories selected as class centers.

Table 9 shows the value of between 8 classes distance as the mean value is higher than 3.4, this shows better clustering results. The results of trajectory clustering in each of the 8 intended classes are shown in Table 10.

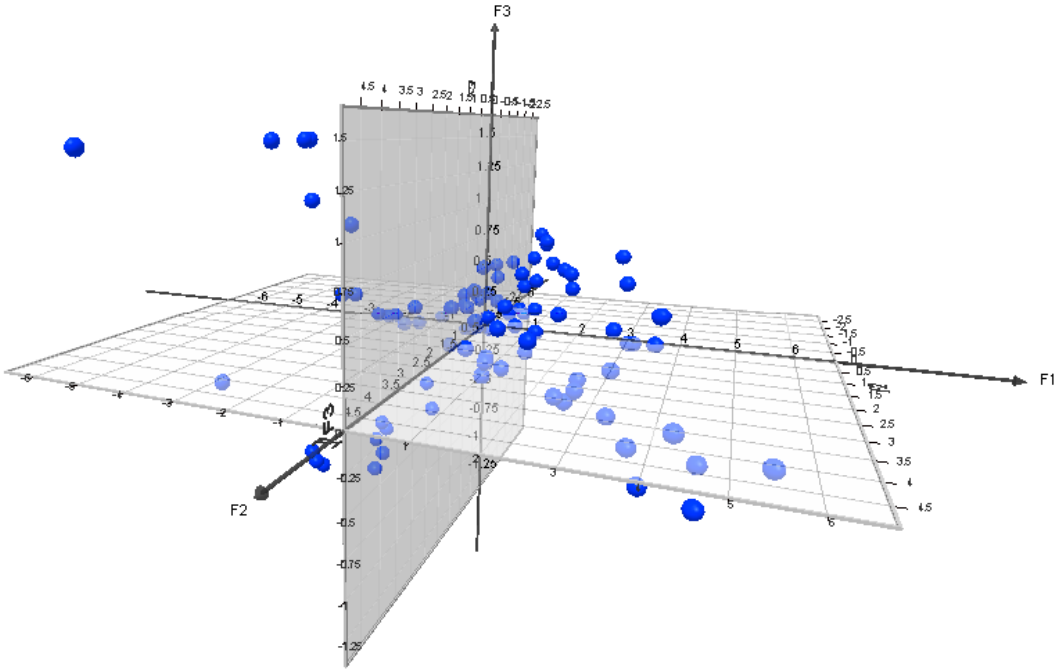


Figure 11. Distributions of observations in the three-dimensional space of the main components

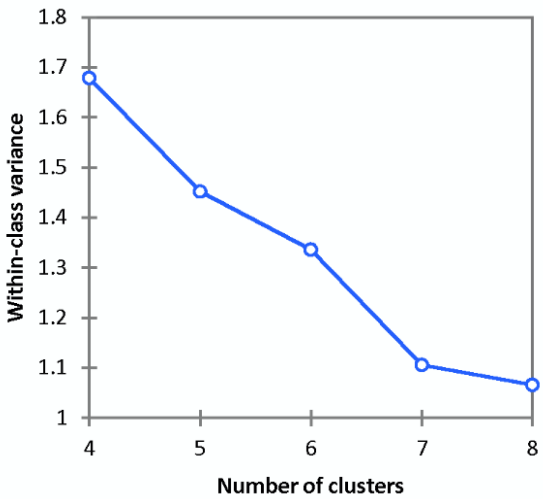


Figure 12. The results of the within-class variance for different clusters

Table 8. Within-class and inter-class variance for different clusters

Variance\Classes	4	5	6	7	8
Within-class	1.682	1.433	1.218	1.095	1.066
Between-classes	5.252	5.501	5.716	5.839	5.868
Total	6.934	6.934	6.934	6.934	6.934

Table 9. The distance between cluster centers

	1 (Obs8)	2 (Obs26)	3 (Obs39)	4 (Obs59)	5 (Obs74)	6 (Obs78)	7 (Obs91)	8 (Obs92)
1 (Obs8)	0	2.783	4.216	3.128	6.036	5.313	6.888	7.946
2 (Obs26)	2.783	0	1.695	1.737	4.187	4.209	5.983	7.131
3 (Obs39)	4.216	1.695	0	1.642	2.738	3.277	4.993	6.100
4 (Obs59)	3.128	1.737	1.642	0	2.923	2.551	4.323	5.457
5 (Obs74)	6.036	4.187	2.738	2.923	0	1.619	2.680	3.644
6 (Obs78)	5.313	4.209	3.277	2.551	1.619	0	1.790	2.923
7 (Obs91)	6.888	5.983	4.993	4.323	2.680	1.790	0	1.163
8 (Obs92)	7.946	7.131	6.100	5.457	3.644	2.923	1.163	0

Table 10. Clustering results for the 8 classes

Class	1	2	3	4	5	6	7	8
Objects	19	24	17	13	6	8	5	2
Sum of weights	19	24	17	13	6	8	5	2
Within-class variance	2.897	0.764	0.418	0.612	0.374	0.464	0.517	0.738
Minimum distance to centroid	0.475	0.366	0.047	0.469	0.306	0.127	0.472	0.607
Average distance to centroid	1.450	0.815	0.569	0.731	0.507	0.579	0.618	0.607
Maximum distance to centroid	4.430	1.363	1.036	1.035	0.921	0.928	0.943	0.607

Table 10 shows that the highest numbers of trajectories are located in the first three classes. Examining these trajectories shows that most of them are trajectories with a medium-length or shorter than 4000 m. The reason for this is the use of the entropy and temporal data criteria of the used parameters, because trajectories with similar lengths are generally temporally similar, although they can vary geometrically.

The results of the clustering for the considered 95 trajectories are shown in Figure 13. Outputs for cluster 2 are illustrated with 24 trajectories presented by red, black, green, and blue colors. Moreover, Figure 12 shows how the mentioned trajectories are similar when considering geometry parameters such as shape and complexity while they could reveal different directions, lengths, and origin-destinations as trajectories presented in Figure 13. Some of these trajectories also share similar paths or distances. Finally, the clustering method might also consider additional parameters, for example, trajectory clusters could be derived by considering origins and destinations.

Overall, the results of the proposed method show 88.66% of general similarity for the trajectories clustered in 8 classes. This precision in similarity is lower for classes 1 to 4 in comparison to other classes, because of the trajectory diversity in these classes. Class 3 has the least similarity between trajectories with a 69.37% value. One solution for improvement in the precision of the aforementioned classes is increasing the number of classes when implementing the K-Means technique while considering all Geolife trajectory data sets. Overall, the proposed approach, by considering critical points and spatio-temporal entropies, provides relatively complete support for clustering and identifying trajectory similarities and movement patterns. As for the Geolife data set (including variety scales and length trajectories), the proposed method reveals similar geometric patterns at different scales and directions. This can be applied to other large trajectory datasets in urban

environments. One peculiarity of the approach is that it integrates geometrical, spatial, and temporal criteria, thus providing a large set of primitives that can be further considered when searching for some specific trajectory patterns. This might provide a relatively large set of options for analyzing trajectory and movement patterns in the city of other contexts.

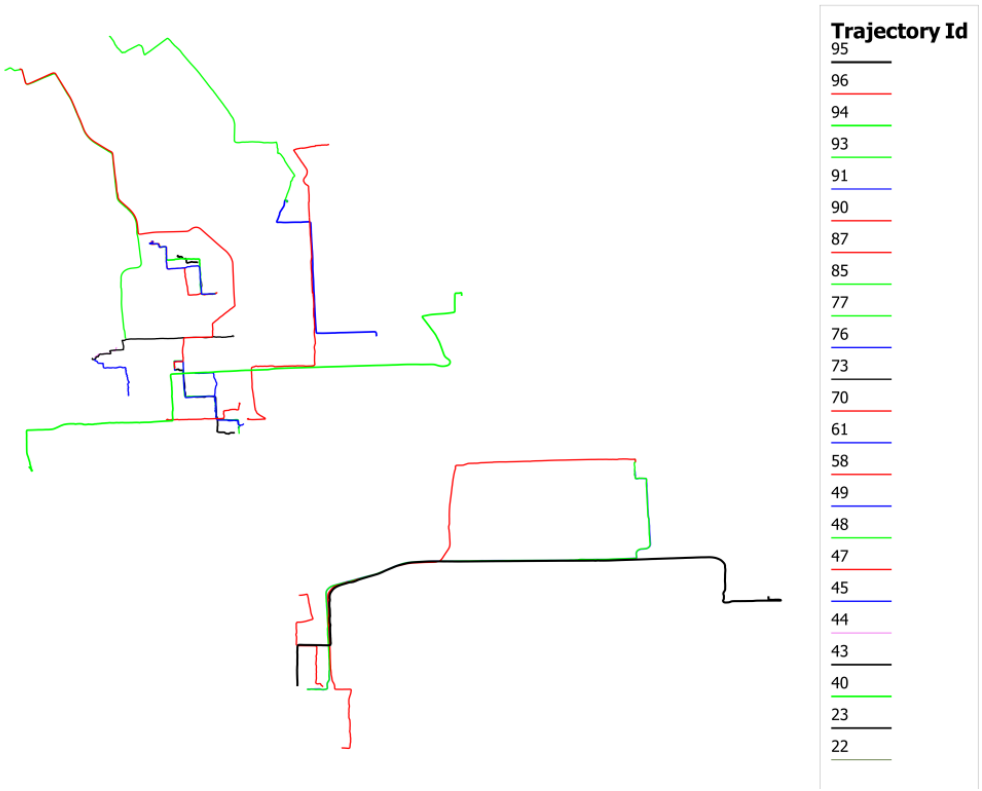


Figure 13. Trajectories of cluster 2

5. Conclusion

Nowadays, with rapid advances in geo-positioning technologies and location-based services, large trajectory datasets are emerging in urban environments. Among the different data widely available, trajectory data provide many opportunities for understanding movement patterns in space and time. Over the past few years, many researchers have paid special attention to identifying geometrically and temporally similar trajectories and obtaining user movement patterns in urban environments. The research presented in this paper develops a computational method to extract similar trajectories. The approach is based on the prior application of the STE-SD method, and that considers both the location and time parameters when studying the similarities of some given trajectories. The main advantages of the proposed method are summarized below:

- One of the main goals is to extract similar trajectories as moving patterns. Several physical and geometric parameters are first considered, and this is according to the spatial and temporal dimensions.
- As the number of potential classes to consider is to identify some clusters, we applied a PCA framework to find the most sufficient components by considering the variance between the different components. This improves the results of the clustering process by decreasing the computational complexity.
- The proposed framework is able to examine the studied trajectories by considering physical or geometrical parameters or their combinations. Also, only temporal and/or spatial or their combination can be considered to compare trajectories.

The proposed method has the potential of identifying geometrically similar trajectories with various directions and durations. Likewise, identifying similar trajectories in the proposed method is regardless of the route's beginning and end point. While the current method is mainly based on geometric and physical parameters, additional spatial, mode of movement, user's specifications, temporal properties should be considered in further work. This is indeed a direction to explore in relation to the semantics (as user's activities) that can be identified from the application domain considered. Finally, while the proposed method is primarily designed and experimented with within the context of urban trajectories, it can be also applied to the

study of migration patterns and animal behaviors at local and regional scales.

References

Asakura, Y., & Hato, E. (2004). Tracking survey for individual travel behaviour using mobile communication instruments. Transportation Research Part C: Emerging Technologies, 12(3), 273-291.

Aung, S. S., & Naing, T. T. (2014). Mining Data for Traffic Detection System Using GPS _enable Mobile Phone in Mobile Cloud Infrastructure. Proceedings of International Journal on Cloud Computing: Services and Architecture (IJCCSA), 4.

Bashir, F. I., Khokhar, A. A., & Schonfeld, D. (2007). Object trajectory-based activity classification and recognition using hidden Markov models. IEEE transactions on Image Processing, 16(7), 1912-1919.

Buchin, M., Dodge, S., & Speckmann, B. (2012). Context-aware similarity of trajectories. International Conference on Geographic Information Science,

Buchin, M., Driemel, A., van Kreveld, M., & Sacristán, V. (2011). Segmenting trajectories: A framework and algorithms using spatiotemporal criteria. Journal of Spatial Information Science, 2011(3), 33-63.

Cao, H., Mamoulis, N., & Cheung, D. W. (2005). Mining frequent spatio-temporal sequential patterns. Data Mining, Fifth IEEE International Conference on,

Demšar, U., Buchin, K., Cagnacci, F., Safi, K., Speckmann, B., Van de Weghe, N., Weiskopf, D., & Weibel, R. (2015). Analysis and visualisation of movement: an interdisciplinary review. Movement ecology, 3(1), 5.

Dodge, S., Weibel, R., & Forootan, E. (2009). Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. Computers, environment and urban systems, 33(6), 419-434.

Dodge, S., Weibel, R., & Laube, P. (2011). Trajectory similarity analysis in movement parameter space. Plymouth, UK: Proceedings of GISRUK, 27-29.

Dodge, S., Weibel, R., & Lautenschütz, A.-K. (2008). Towards a taxonomy of movement patterns. Information visualization, 7(3-4), 240-252.

El Mahrsi, M. K., & Rossi, F. (2012). Graph-based approaches to clustering network-constrained trajectory data. International Workshop on New Frontiers in Mining Complex Patterns,

Fang, H., Hsu, W.-J., & Rudolph, L. (2009). Mining user position log for construction of personalized activity map. International Conference on Advanced Data Mining and Applications,

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. Transactions in GIS, 17(3), 463-481.

Giannotti, F., & Pedreschi, D. (2008). Mobility, data mining and privacy: Geographic knowledge discovery. Springer Science & Business Media.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779-782.

Harguess, J., & Aggarwal, J. (2009). Semantic labeling of track events using time series segmentation and shape analysis. 2009 16th IEEE International Conference on Image Processing (ICIP),

Himberg, J., Korpiaho, K., Mannila, H., Tikanmaki, J., & Toivonen, H. T. (2001). Time series segmentation for context recognition in mobile devices. Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on,

Hofmann, M., Wilson, S. P., & White, P. (2009). Automated identification of linked trips at trip level using electronic fare collection data. Transportation Research Board 88th Annual Meeting,

Hornsby, K. S., & Cole, S. (2007). Modeling Moving Geospatial Objects from an Event-based Perspective. Transactions in GIS, 11(4), 555-573.

Hosseinpoor Milaghardan, A., Ali Abbaspour, R., & Claramunt, C. (2018a). A Geometric Framework for Detection of Critical Points in a Trajectory Using Convex Hulls. ISPRS International Journal of Geo-Information, 7(1), 14.

Hosseinpoor Milaghardan, A., Ali Abbaspour, R., & Claramunt, C. (2018b). A Spatio-Temporal Entropy-based Framework for the Detection of Trajectories Similarity. Entropy, 20(7), 490.

Jolliffe, I. (2011). Principal component analysis. In International encyclopedia of statistical science (pp. 1094-1096). Springer.

Kafkafi, N., & Elmer, G. (2005). Texture of locomotor path: a replicable characterization of a complex behavioral phenotype. Genes, Brain and Behavior, 4(7), 431-443.

Kafkafi, N., Yekutieli, D., & Elmer, G. I. (2009). A data mining approach to in vivo classification of psychopharmacological drugs. Neuropsychopharmacology, 34(3), 607-623.

Laube, P., & Purves, R. S. (2011). How fast is a cow? cross-scale analysis of movement data. Transactions in GIS, 15(3), 401-418.

Lee, J.-G., Han, J., Li, X., & Cheng, H. (2011). Mining discriminative patterns for classifying trajectories on road networks. IEEE Transactions on Knowledge and Data Engineering, 23(5), 713-726.

Lee, J.-G., Han, J., Li, X., & Gonzalez, H. (2008). TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering. Proceedings of the VLDB Endowment, 1(1), 1081-1094.

LIN, H., LV, J., YANG, C., DENG, M., WANG, K., & WANG, X. (2014). GPS Trajectory Mining: a Survey. Journal of Computational Information Systems, 10(16), 6947-6956.

Lin, M., & Hsu, W.-J. (2014). Mining GPS data for mobility patterns: A survey. Pervasive and Mobile Computing, 12, 1-16.

Lu, M., Wang, Z., Liang, J., & Yuan, X. (2015). OD-Wheel: Visual design to explore OD patterns of a

central region. Visualization Symposium (PacificVis), 2015 IEEE Pacific,

Milaghardan, A. H., Abbaspour, R. A., & Claramunt, C. (2018). A Dempster-Shafer based approach to the detection of trajectory stop points. Computers, environment and urban systems.

Miller, H. J., & Han, J. (2009). Geographic data mining and knowledge discovery. CRC Press.

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009). Wherenext: a location predictor on trajectory pattern mining. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining,

Morzy, M. (2007). Mining frequent trajectories of moving objects for location prediction. Machine Learning and Data Mining in Pattern Recognition, 667-680.

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., & Pelekis, N. (2013). Semantic trajectories modeling and analysis. ACM Computing Surveys (CSUR), 45(4), 42.

Pelekis, N., Kopanakis, I., Kotsifakos, E., Frentzos, E., & Theodoridis, Y. (2009). Clustering trajectories of moving objects in an uncertain world. Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on,

Perttunen, M., Kostakos, V., Riekki, J., & Ojala, T. (2015). Urban traffic analysis through multi-modal sensing. Personal and Ubiquitous Computing, 19(3-4), 709-721.

Robinson, A. C., Peuquet, D. J., Pezanowski, S., Hardisty, F. A., & Swedberg, B. (2017). Design and evaluation of a geovisual analytics system for uncovering patterns in spatio-temporal event data. Cartography and Geographic Information Science, 44(3), 216-228.

Soleymani, A., Cachat, J., Robinson, K., Dodge, S., Kalueff, A., & Weibel, R. (2014). Integrating cross-scale analysis in the spatial and temporal domains for classification of behavioral movement. Journal of Spatial Information Science, 2014(8), 1-25.

Zheng, Y., Chen, Y., Li, Q., Xie, X., & Ma, W.-Y. (2010). Understanding transportation modes based on GPS data for web applications. ACM Transactions on the Web (TWEB), 4(1), 1.

Zheng, Y., Zhang, L., Xie, X., & Ma, W.-Y. (2009). Mining interesting locations and travel sequences from GPS trajectories. Proceedings of the 18th international conference on World wide web,

Zhou, Y., Fang, Z., Thill, J.-C., Li, Q., & Li, Y. (2015). Functionally critical locations in an urban transportation network: Identification and space–time analysis using taxi trajectories. Computers, environment and urban systems, 52, 34-47.