# A Context-Aware Dimension Reduction Framework for Trajectory and Health Signal Analyses

Samira Goudarzi [1], Mohammad Sharif [3] and Farid Karimipour [1,2]

[1] *School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran*

[2] *Institute of Science and Technology (IST Austria), Klosterneuburg, Austria*

[3] *Department of Geography, Faculty of Humanities, University of Hormozgan, Iran*



## Abstract

It is practical to collect a huge amount of movement data and environmental context information along with the health signals of individuals because there is the emergence of new generations of positioning and tracking technologies and rapid advancements of health sensors. The study of the relations between these datasets and their sequence similarity analysis is of interest to many applications such as health monitoring and recommender systems. However, entering all movement parameters and health signals can lead to complexity of the problem and to an increase in its computational load. In this situation, dimension reduction techniques can be used in order to avoid consideration of simultaneous dependent parameters in the process of similarity measurement of the trajectories. The present study provides such a framework to use spatial-temporal data and movement parameters along with independent context information in the process of measuring the similarity of trajectories. In this regard, the omission of dependent movement characteristic signals is conducted by using an unsupervised feature selection dimension reduction technique. To evaluate the effectiveness of the proposed framework, it was applied to a real contextualized movement and related health signal datasets of individuals. The results indicated the capability of the proposed framework in measuring the similarity and in decreasing the characteristic signals in such a way that the similarity results - before and after reduction of dependent characteristic signals- have small differences. The mean differences between the obtained results before and after reducing the dimension were 0.029 and 0.023 for the round path, respectively.

**Keywords** Movement, Context-Aware, Signal Characteristic, Similarity Measurement, Dimension Reduction, Health GIS

## 1 Introduction

Movement is a complex multidimensional process involving space, time, and context (Dodge, 2019). Moving entities (e.g., humans, animals, and vehicles) leave a trace of their position over time, which is referred to as *spatial trajectories* (Zheng and Zhou, 2011). In recent years, due to significant advancements in tracking and navigation technologies (e.g., Global Positioning System (GPS)) a large amount of movement data is recorded in the form of spatial trajectories along with various temporal resolutions (Basiri et al., 2018). In addition, it is possible to ubiquitously monitor the environment and collect contextual information (e.g., weather conditions, slope, traffic, and health signals) over time by fixed and moveable sensors along with spatial trajectories (Kays et al., 2015, Sharif and Sadeghi-Niaraki, 2017), which we referred to as *contextual trajectories*. For example, some wearable gadgets are equipped with sensors to measure the health signs of individuals instantly (like heart rate and blood pressure), along with the capability of recording the position and movement parameters, such as speed and direction of individuals by their embedded navigation systems (e.g., GPS). Therefore, studying behaviors of moving entities, while considering various context factors, challenges the existing methods of data mining and makes their development more essential than ever.

Similarity analysis of trajectories and movement signals processing are fundamental issues in spatial data mining. It can be considered as a low-level knowledge extraction technique because the similarity outputs can be used in other data mining techniques, such as movement pattern recognition, classification, and clustering. On the one hand, finding the similarities between trajectories have been conducted by traditional geometrical functions that neglect the inherent nature of the entities. On the other hand, a large number of studies have used spatial, temporal, and movement parameters in similarity analysis of trajectories. To have a more accurate understanding of movement patterns of entities in different conditions, it is necessary to establish a relation between the movement of an entity and its context (Amouzandeh et al., 2018, Brum-Bastos et al., 2018, Sharif et al., 2019). Therefore, for context-aware analysis of trajectories there is a need to develop algorithms that can compare the context information along with spatial parameters and movement sequences of entities (Sharif et al., 2018). Indeed, analyzing all the aforementioned parameters simultaneously may lead to more complexity of the problem and an increase in the algorithm's computational load. One possible solution is to find out the interdependent trajectories variables and disregard them in the similarity analysis process. This can be accomplished by developing

dimension reduction methods and by only entering the maximum independent movement data and context information.

Context-aware movement analysis in the health sector has become more noticeable than ever. Monitoring individuals' health by analyzing their functional signals during movement is considered as one of the significant cases in this context (Kaffash-Charandabi et al., 2019). It is desired to study the relation of human health indexes (e.g., heart rate and blood pressure), movement parameters (e.g., speed, direction, and acceleration), and environmental contexts (e.g., slope and elevation) by measuring the context-aware similarity of trajectories. However, in ubiquitous monitoring of individuals' health, the significance of fast processing of the generated big movement data and context information is essential. Therefore, discovering and reducing dependent parameters for faster computation is possible by using the dimension reduction techniques in the study of both trajectories and functional signals.

Previous researches on trajectory similarity analysis have considered all the attributes of trajectories simultaneously. That causes complexity and computational load. In addition, previous researches have employed geometrical functions to compare the distances between these attributes- which have their own limitations. Therefore, this research aims to present a context-aware framework for similarity analysis of trajectories, which removes the dependent movement and context information by using a dimension reduction technique. The contributions of this study are mainly: (1) finding the effective variables of spatial, movement, and context (environmental and individual) as well as their inter-relation (interaction) in the analysis of individual's movement and health data, and (2) providing a framework based on dimension reduction to measure the similarities of trajectories and health signals of individuals by maintaining the value of similarities. The results of this study can be used to predict the future trend of movement and its effects on the health signs of individuals in different systems including monitoring and health recommender systems.

The remainder of this article is structured as follows: Section 2 explains the basic concepts and a review of previous research on similarity measurement of trajectories. Section 3 proposes a context-aware similarity measurement framework based on dimension reduction techniques. Section 4 evaluates the developed framework by its implementation on the actual data of pedestrians. The obtained results are discussed in Section 5. Finally, conclusions and suggestions for future research are discussed in Section 6.

## 2 Preliminary and Review of Literature

This section begins by introducing the concepts used in conducting the objectives of this research. The following is a review of the research conducted in the field of movement analysis based on the similarity measurements of the trajectories using spatial data, movement parameters, and context information. Finally, it introduces the existing methods and conducted research in dimension reduction including feature selection and extraction.

### 2.1 Concepts

#### *2.1.1 Similarity measure of trajectories*

Basic analysis of trajectories is a similarity analysis that answers the question: "How much are the trajectories of two or more entities similar?" Trajectories can be considered similar if they have the same shapes and movement parameters, visit the same places, have identical patterns (Laube, 2014), and share commonalities in their contexts (Sharif and Alesheikh, 2017). In general, the methods of determining the similarity of trajectories can be divided into three categories of geometric similarity based on space and time, physical similarity based on movement properties, and context-aware similarity based on the composition of properties (Demšar et al., 2015). In addition, several methods investigate the similarity of two signals based on the shapes and behaviors of the signals rather than their distance. These methods are normally used in the fields of artificial intelligence, machine learning, and pattern recognition and classification (Chen et al., 2009, Duch, 2000) as well as in GIScience and remote sensing applications, to measure spectral properties between two different vectors. Examples of these applications and research include dimension reduction (Hasanlou et al., 2015), target recognition (Zhang et al., 2015), change detection (Liu et al., 2015), image registration (Ren et al., 2016), image classification (Hosseini et al., 2012), and embedding representation (Nalmpantis and Vrakas, 2019). Among the shape-based similarity measure methods, Kulczynski (KU), Taminoto (TA) and Pearson correlation coefficient (PCC) have been of interest (Seydi and Hasanlou, 2017). In addition, PCC and their variants have been widely used in collaborative filtering to evaluate similarity (Wang et al., 2017). In a separate study, the performance, strengths, and weaknesses of similarity measurement methods were reviewed (Ranacher and Tzavella, 2014).

### *2.1.2 Dimension reduction*

Feature selection and feature extraction methods have been introduced for dimension reduction. In feature selection method, the desired output is a subset of initial data. In fact, a subset of the original data, which is smaller than the initial set, is chosen without any change in the selected data. However, in feature extraction method, the nature of data changes and subsequently the data computing space changes (Krızek, 2008). Therefore, one of these methods is selected depending on the application and data types. Feature selection method includes supervised filtering techniques and wrapper, and unsupervised similarity-based, mutual information, and dependency-based techniques. To achieve the research objective, the feature selection method is used, since the nature of the signals will be preserved after reducing dimension and selecting the optimal and efficient signals.

## 2.2 Review of Literature

A large number of studies have used distance-based similarity functions for spatial and spatial-temporal similarity measurements (Yuan and Raubal, 2014, Xia et al., 2010). Given the variety of movement properties along with spatial-temporal information, Dodge et al. (2012) generated trajectories based on movement parameters (i.e., speed, acceleration, and direction) and found the similar trajectories by applying a developed version of the editing distance, called the normalized weighted edit distance (NWED). Adding semantic information to the trajectories makes the questions and answers, analysis, and data mining of moving objects' data simpler (Alvares et al., 2007). Furtado et al. (2016) transformed the trajectories of moving objects into a sequence of stops and considered a set of features (the purpose of stop and activity at stop position) which resulted in a multi-dimensional trajectory.

Movement in the real world is significantly affected by both internal and external contexts. The term *context* in the movement studies is defined as "that part of a situation or data that influences movement or is influenced by movement" and has been divided into two forms of internal context (i.e., all details directly related to the entity) and external context (i.e., any environmental factor that affects the movement) (Sharif and Alesheikh, 2018). Due to the importance of context in the movement behavior analysis, various studies have been conducted in this field including ranking and reduction of uncertainty of trajectories based on context information (Dai et al., 2016), prediction of the user's external location based on context information (Guessoum et al., 2016), and human movement analysis using a combination of trajectories made by GPS and context information (Siła-Nowicka et al.,

2016). Similarity measures of trajectories may undergo changes while considering the context dimension. Buchin et al. (2014) presented a similarity measurement framework by combining context and spatial distances. These distance functions were the pre-defined Fréchet distance, Hausdorff distance, and equal time distance. Sharif and Alesheikh (2017) provided a distance measurement function called context-based dynamic time warping (CDTW) by using context information in similarity search of aircraft trajectories in order to find spatial-temporal-context patterns in their databases. Context information has its own heterogeneity and uncertainty because it is collected from different sources and with quantitative (numerical) and qualitative (descriptive) values. Therefore, Sharif et al. (2019) developed a context-aware framework based on fuzzy inference systems to model the similarity of trajectories using different types of movement data and internal and external context information. A review of the relevant literature reveals that almost every research has exploited traditional similarity measure functions that are computationally overpriced. In addition, every research considered all the attributes of trajectories that caused a high computational load.

Dimension reduction methods have been suggested to reduce computational costs especially in hyperspectral image analysis. Zhao and Du (2016) presented a framework based on spectral-spatial band feature-based classification (SSFC) using dimension reduction and deep learning techniques. Ren et al. (2014) used the dimension reduction principal component analysis (PCA) procedure in hyperspectral band imagery to overcome large dimensions handling. In the data mining field, Houari et al. (2016) proposed a new technique for reducing the high-dimensional data using the Copulas and Lu analysis tools because the use of high-dimensional data increases noise, redundant data, and the possibility of unconnected data entities. General feature selection techniques are developed based on mutual information using single-objective and multi-objective optimizations. To solve the problem of discrete and continuous random variables in a dataset, Coelho et al. (2016) presented a mutual information estimator to solve the problems of dimension reduction and classification. Given the good functionality of dimension reduction methods, they are presumed to be feasible to be employed in movement analysis in terms of detecting and removing dependent context information.

By reviewing previous researches, it can be concluded that it is possible to simultaneously apply spatial, movement, and contextual dimensions in the similarity measure process. However, realizing the interactions between these parameters and removing the dependent parameters may reduce the computation time. Besides, moving objects' trajectories have been

analyzed by traditional distance-based functions in a pairwise manner where each function has its limitation and complexity. As an alternative, previous researches have shown the robustness of shape-based methods in signal analysis, which can be applied to finding the similarities of trajectories. Therefore, the proposed framework in this research uses shape-based measurement along with only optimum context dimensions. It may be employed in applications -like recommender systems- that require high-speed decision-making.

## 3 Methodology

The methodology of the research consists of three main stages, shown graphically in Fig. 1: (1) data preparation, (2) unsupervised feature selection using mutual information, and (3) evaluation of the mutual information method performance by computing the similarity before and after dimension reduction.

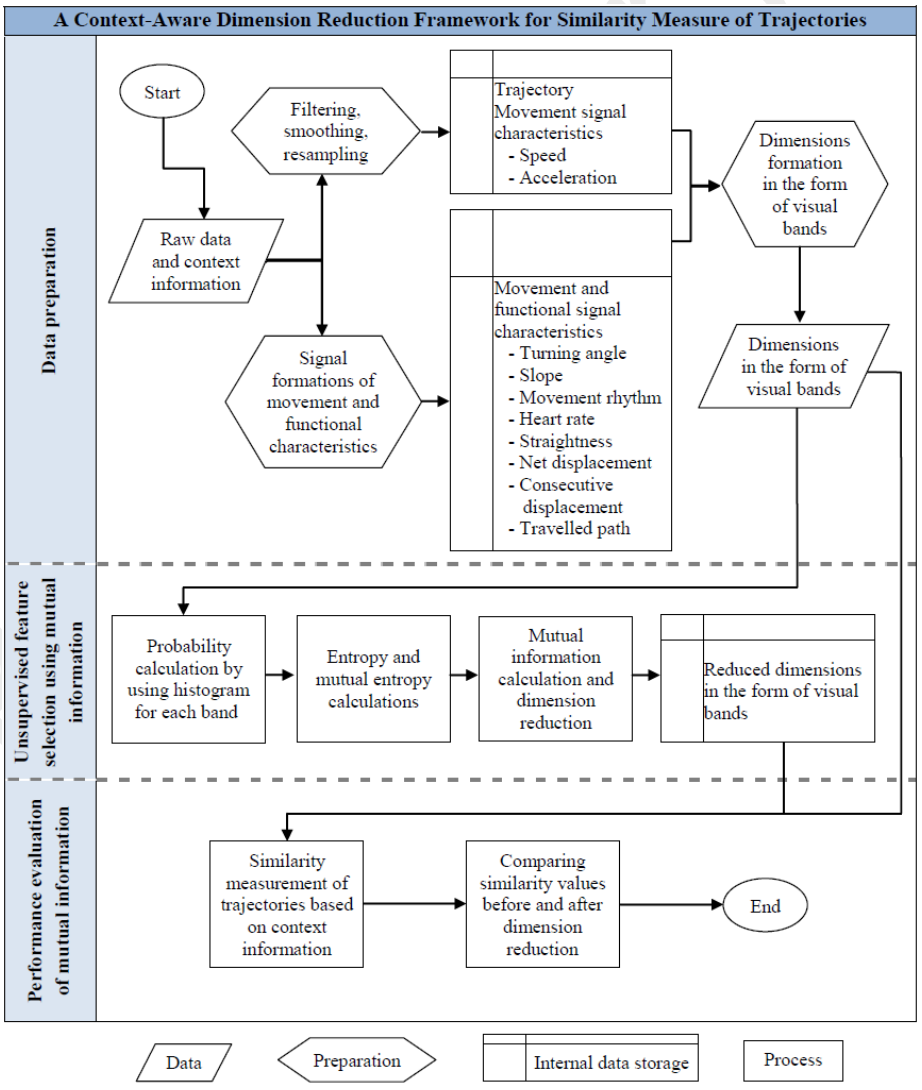



**Fig. 1** Methodology for analyzing movement signals of individuals

## 3.1 Data preparation

### *3.1.1 Filtering, smoothing, and resampling*

The precision of GPS raw data depends on the presence of obstacles which block GPS signals, multi-path effects, ionospheric and tropospheric errors, and so on. Preprocessing of trajectories to detect, correct, and remove erroneous records to increase the quality of the trajectories is fundamental before the movement data analysis (Yan and Spaccapietra, 2009). The data preprocessing step normally consists of filtering, resampling, smoothing, and/or map matching.

Filtering is used as a statistical approach to detect and remove outlier(s). It is considered as an outlier if any point of the trajectory is distanced from its predecessor point as 2.5 times as the standard deviation of the distance between the consecutive points of the trajectory. Furthermore, linear interpolation techniques are used in the resampling step to create spatial trajectories with regular sampling intervals. The smoothing process is used to remove the noise effect of data. More is used in navigation applications and transportation networks, and there is no need for a map-matching step in this research. Various methods including least square, Spline approximation, moving average, Kernel-based methods, and Kalman filter have been used to smooth the raw GPS data (Dodge et al., 2009). Since the data of this research is collected by the Garmin device with defined precision, the extended Kalman filter is used to smooth and filter the data (Jun et al., 2006). In addition, linear interpolation is used in resampling. The output of the Kalman filter is spatial trajectories and movement parameter signals such as speed and acceleration.

### *3.1.2 Formation of the characteristic signal*

Studying entities' trajectories spatially along with their movement characteristics (e.g., speed, acceleration, and direction) and their relations have always been of interest (Giannotti and Pedreschi, 2008). As movement starts, a continuous time-dependent signal is created due to the change in movement characteristics, which is referred to as *movement parameter signal* (MPP) (Dodge et al., 2009). These parameter signals can be in the forms of spatial (such as x, y, z coordinates), movement (such as speed, acceleration of the individual), and performance (such as heart rate). Contrary to the spatial trajectories indicating the geometric characteristics of the movement over time, the movement characteristics represent the physical concepts of movement. Therefore, these characteristics describe the movement of the entity and help to understand movement behaviors more comprehensively, and are also

used in the exploration of the behavioral patterns of moving objects (Nathan et al., 2008). Detecting different modes of movement such as walking, bikes, cars, and buses, or ecological behaviors of animals including standing, flying, and forage are easily identifiable by dividing the trajectory using quantitative parameters (Laube, 2014). In this research, the movement parameter signals include speed (the displacement rate of the entity), acceleration (the rate of speed change of the entity), and the Kalman filter output. To enter the process of similarity measuring, the movement parameter signals include net displacement (beeline distance between a point and the initial point of the profile), consecutive displacement (beeline distance between two consecutive points), turning angle (direction of movement), straightness (path rate and consecutive displacement), traveled path (length of the trajectory), and slope (height differences and distance differences) are calculated. Furthermore, the movement parameter signal of cadence (number of steps per minute) and the performance parameter signal of heart rate are collected by a device.

### *3.1.3 Formation of dimensions for all moving entities in the form of visual bands*

To implement the unsupervised feature selection of mutual information, all spatial trajectories and characteristic signals should be converted into visual bands. The O symbol is considered for a moving entity and the associated dataset is defined as $O^i(X_1^i, Y_2^i, Z_3^i, PP_4^i, PP_5^i, ..., PP_m^i)$, where m represents the number of signals collected for the $i^{th}$ moving entity. As shown in Fig. 2, the construction of the dimension of one of the parameter signals is indicated for all moving entities. $(PP_1^i, ..., PP_n^i)$ represents the vector of one of the parameter signals for the $i^{th}$ moving object, where n is the magnitude of the vector.

| | | | | | |
|---|---|---|---|---|---|
| $PP_1^1$ | $PP_2^1$ | . | . | . | $PP_n^1$ |
| $PP_1^2$ | $PP_2^2$ | . | . | . | $PP_n^2$ |
| . | . | | | | . |
| . | . | | | | . |
| . | . | | | | . |
| $PP_1^i$ | $PP_2^i$ | . | . | . | $PP_n^i$ |

**Fig. 2** The process of generating dimension for a characteristic signal for every moving entity

### 3.2 Unsupervised Feature Selection using Mutual Information

The mutual information method, which is regarded as one of the unsupervised feature selecting methods, is used in this research. Mutual information is a complete statistic for measuring the dependence between two random variables (Cover and Thomas, 2012). Furthermore, mutual information can be easily interpreted, while the interpretation of other statistics such as rank-ordering is more difficult (Ross, 2014). Another definition of mutual information is the reduction of uncertainty due to another random variable (Cover and Thomas, 2012). The Shannon Information Unit (symbol: sh), which is known as a bit, is defined by IEC 80000-13. A Shannon is the content information of an event whose occurrence probability is 0.5 (Stone, 2015). Mutual information for discrete (Cover and Thomas, 2012), continuous (Stone, 2015), and both discrete and continuous random variables can be calculated (Ross, 2014). The calculation method of mutual information for a visual band is described as follows since all the spatial trajectories and parameter signals are converted into visual bands. Suppose the image X is available with N pixels. The first thing to do for calculating the entropy and mutual information is to partition ($X_1$, $X_2$, ..., $X_i$, ..., $X_t$) to show that each pixel is assigned to a group based on its value. $X_i$ is the number of pixels belonging to the $i^{th}$ group. ($P_1$, $P_2$, ..., $P_i$, ..., $P_t$) are the occurrence probability of the values in the image. In other words, the probability values can be obtained using a histogram of an image.

The Shannon entropy represents the amount of information in a visual band. Shannon entropy of an image is calculated by using Eq. 1 (Wang et al., 2012).

$$H(X) = -E[\log p(x)] = -\sum_{x \in X} p(x) \log(p(x)) \quad (1)$$

The mutual information of two visual bands is calculated by Eq. 2, which indicates the amount of information that is mutual in two visual bands (Wang et al., 2012).

$$I(X;Y) = H(X) - H(X \mid Y) \quad (2)$$

In terms of mutual information, H(X) and H(Y) are Shannon entropies for the two X and Y bands. H(X, Y) is the conditional entropy for the two X and Y bands. The obtained value from the mutual information is not between 0 and 1 range, but the obtained result from the mutual information is now standing within the unit interval because of using the normal mutual information,. Eq. 3 indicates the computation method of mutual information.

$$I_{normal}(X,Y) = 2.\left(\frac{I(X,Y)}{H(X)+H(Y)}\right) \quad (3)$$

When the obtained value is closer to one (1), the shared information in two bands becomes more and as the value is closer to zero (0), the shared information by the two bands becomes less.

A threshold is defined in using mutual information techniques to remove the dependent dimensions. If the mutual information of two dimensions is greater than the threshold, they will be kept in the dataset. However, one of these two dimensions should be removed if mutual information of the two dimensions is greater than the threshold value. The way of choosing the threshold is empirical. More details on selection procedure of the threshold are explained in the implementation section.

### 3.3 Performance Evaluation of the Proposed Framework

To evaluate the performance of the mutual information method, the comparison of the obtained results from the context-based similarity measurement process is required before and after the dimension reduction. First, all the signal parameters and spatial trajectories are entered into the similarity measurement process and the similarity value is obtained for both moving entities. Then, the dimensions are removed through using the mutual information technique. Finally, the reduced dataset enters the context-based similarity measurement process. Two series of similarity values are compared before and after dimension reduction in order to evaluate the proposed approach of the research. The context-based similarity measurement process is explained below.

#### *3.3.1 Context-based similarity measurement process*

In this research, a method for measuring the similarity of two moving entities is introduced. This method can enter all context dimensions in the process of similarity measurement and it has a higher speed in comparison to other methods. Furthermore, this method can operate more efficiently when there is context information in the dataset collected for a moving entity. Using this method, the simultaneous effect of two dependent dimensions can be avoided in terms of the value obtained for the similarity value. The context-based similarity measurement framework for moving entities is explained as follows.

Letter O represents a moving entity. The associated dataset is defined as $O^i(C^i_{1,1}, C^i_{1,2}, ..., C^i_{1,m_1}, C^i_{2,1}, C^i_{2,2}, ..., C^i_{2,m_2}, ..., C^i_{n,m_n})$. In addition, $\sum_{K=1}^{n} m_K$ equals the number of signals collected for the $i^{th}$ moving entity. The context-based similarity measurement approach includes the following four steps:

**Step 1**: The visual bands are created based on the number of signals collected for each moving object. The number of created visual bands is $\sum_{K=1}^{n} m_K$. If the existing bands are categorized in such a way that the dependent bands are gathered in one category, n is the number of categories, and $m_k$ is the number of dependent bands belonging to each category.

**Step 2**: Dependent dimensions for moving entities were determined in the first step. The second step is to consider two moving entities to calculate the similarity of each band in each category. As it was already mentioned, there are methods that measure the similarity of two signals based on the shape and behavior of the signals rather than on the distance between them. In these shape-based methods, signals can be used to assess the similarity of spatial trajectories and functional and movement characteristic signals. In this research, one of the shape-based similarity measurement methods of signals is used because the obtained similarity value for spatial trajectories and functional and movement characteristic signals with different numerical scales are in the same range. In addition, this method is faster than the distance-based methods.

One of the similarity measurement methods based on the shape of the signal is the *Pearson correlation coefficient* (PCC). This method expresses the dependence of two signals. The more similar the behavior and the shape of the two signals are, the greater dependence the two signals indicate. Therefore, this method can be considered as an appropriate criterion to measure the similarity of two signals. The computational formula of this method is presented for two signals of $C^i_{n,m_n}$ and $C^j_{n,m_n}$ in Eq. 4 (Wang, 2012),

$$S^{i,j}_{C_{n,m_n}} = \frac{\sum_{i=1}^{n}\left(C^i_{n,m_n} - \mu_{C^i_{n,m_n}}\right).\left(C^j_{n,m_n} - \mu_{C^j_{n,m_n}}\right)}{\left(\sum_{i=1}^{n}\left(C^i_{n,m_n} - \mu_{C^i_{n,m_n}}\right)\right)^{0.5} \cdot \left(\sum_{i=1}^{n}\left(C^j_{n,m_n} - \mu_{C^j_{n,m_n}}\right)\right)^{0.5}} \quad (4)$$

where $\mu_{C^{i}_{n,m_n}}$ indicates the mean of the values of signal $C^{i}_{n,m_n}$ and $\mu_{C^{j}_{n,m_n}}$ represents the mean of the values of the signal $C^{j}_{n,m_n}$. The $S^{i,j}_{C_{n,m_n}}$ is the obtained similarity for the signal $C^{i}_{n,m_n}$ and $C^{j}_{n,m_n}$ for the i and j entities. Therefore, the similarity is calculated by using Eq. (4) in each category for each dimension.

**Step 3:** The average is obtained according to Eq. (5) from the obtained similarity values in each category.

$$S^{i,j}_{n} = \frac{S^{i,j}_{C_{n,1}} + ... + S^{i,j}_{C_{n,m_n}}}{m_n} \tag{5}$$

where $S^{i,j}_{n}$ is the obtained similarity in the $n^{th}$ category. Therefore, a similarity value is obtained for each category. The n values of similarity are computed because the number of categories is equal to n.

**Step 4:** Finally, an average is derived from n similarity value. The computational similarity formula for the moving objects $O_i$ and $O_j$ is according to Eq. 6.

$$\begin{aligned} S^{i,j} &= \frac{\frac{S^{i,j}_{C_{1,1}} + ... + S^{i,j}_{C_{1,m_1}}}{m_1} + \frac{S^{i,j}_{C_{2,1}} + ... + S^{i,j}_{C_{2,m_2}}}{m_2} + .... + \frac{S^{i,j}_{C_{n,1}} + ... + S^{i,j}_{C_{n,m_n}}}{m_n}}{n} \\ &= \frac{S^{i,j}_{1} + S^{i,j}_{2} + ... + S^{i,j}_{n}}{n} \end{aligned} \tag{6}$$

where $S^{i,j}_{C_{n,m_n}}$ is considered as the similarity obtained for $C^{i}_{n,m_n}$ and $C^{j}_{n,m_n}$ signals using the PCC. $S^{i,j}$ is the final similarity obtained for i and j moving entities.

## 4 Implementation and Results

This section firstly introduces the study area and dataset and then describes the effective movement and context parameters attributed to individuals' health. Secondly, it evaluates the proposed framework and demonstrates its efficiency in reducing the number of collected movement signals by implementing it on a movement dataset.

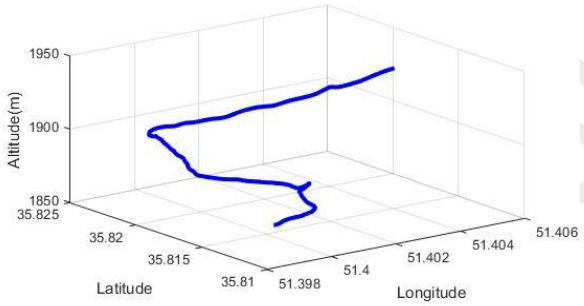
### 4.1 Data

The data were collected in a region in the north of Tehran, Iran, called Bam-e-Tehran. Bame Tehran refers to the upper part of Velenjak region. Bam-e-Tehran complex starts at an altitude of 1830 meters in Velengak Street and ends at an altitude of 1920 meters above sea level. This 2-km long and high-slope path is used as a recreational and pedestrian path for its beautiful mountain views in the highest point of Tehran city (Fig. 3a).

The dataset comprises of the time series of spatial, movement, and functional characteristics of individuals' movements. In particular, 50 trajectories of 25 different pedestrians are collected so that each person traveled (going up and returning down) the designated path on foot. Hereinafter, we name the upward and downward paths as path 1 and path 2, respectively. The traveled path and its profile are illustrated in Fig. 3b. The *spatial data* including latitude, longitude, and altitude, the *movement signals* including speed, acceleration, net displacement, consecutive displacement, turning angle, straightness, and traveled path, the *functional signals* including heart rate, and cadence (the number of steps per minute), and path *slope* (the height difference and distance difference) of each moving object are either collected or calculated. The data collection rate is 1 second. The averages of the total sampling points for all trajectories and the length of path are, respectively, 1303 and 0.991 km for path 1, and 1323 and 0.981 km for path 2. The spatial trajectory units including x, y, height, net displacement, consecutive displacement, and traveled path are in meter. Furthermore, units of speed, acceleration, turning angle, and heart rate are m/s, $m/s^2$, radians, and beats per minute (BPM), respectively.

The employed device for data acquisition is a Garmin GPS smart watch, Vivoactive HR model. Among its several features, the Elevate™ wrist heart rate technology enables us to measure heart rate from the wrist. Furthermore, using its built-in GPS, the Garmin device allows us to have spatial data including latitude, longitude, and altitude synchronized with time. It also calculates speed, acceleration, slope, distance, time, and cadence. This wearable device can be used with the web-based Garmin Express software.


(a)




(b)

**Fig. 3** Study region: (a) the traveled path by individuals, (b) path profile in 3D

As already mentioned, the GPS data needs to be preprocessed. The extended Kalman filter algorithm is implemented on latitude, longitude, and height data. The outputs of the Kalman filter are 3D spatial trajectories (x, y, and height), which are smoothed and filtered. In addition, speed and acceleration movement signals were calculated using the Kalman filter. Further, the collected and calculated signals were resampled to make them symmetric.

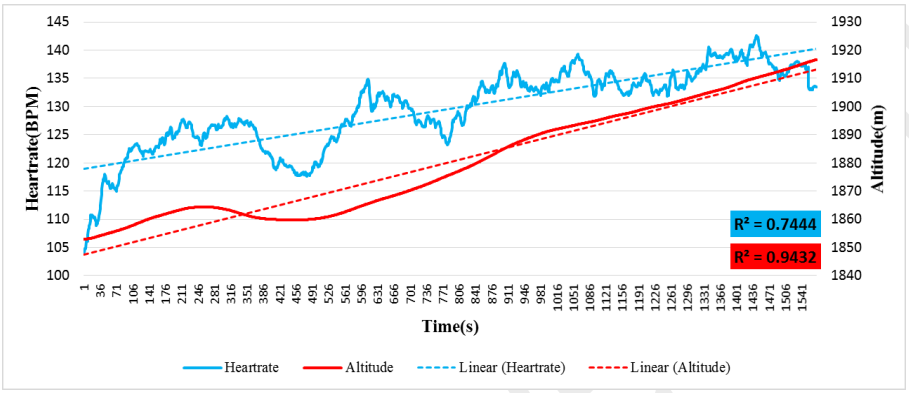
## 4.2 Results

The proposed research model was implemented on the dataset. In the following, the implementation results are presented in two sub-sections. The first sub-section examines the relations between movement parameters. The second examines the performance of the proposed framework of the research and clarifies in detail the results of the evaluation of similarity values before and after the feature selection procedure.

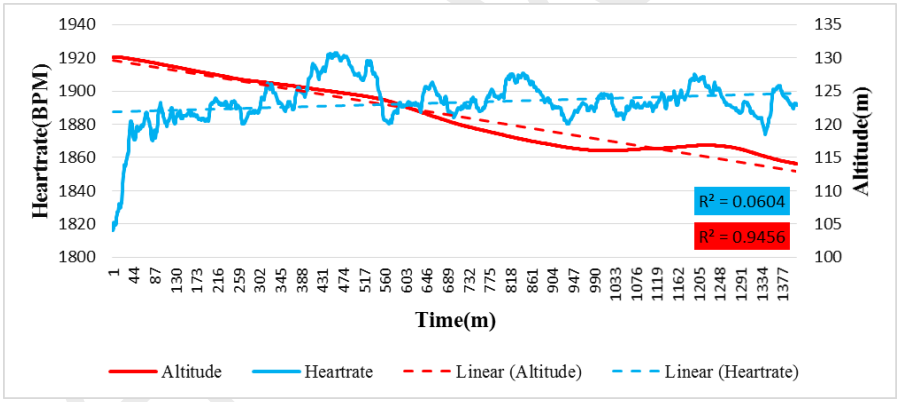
### *4.2.1 Investigating the relation between movement parameters*

This section examines the relations in the dataset. There certainly are relations between for the round trip path, the signals of paths, the height and the heart rate, the height and the

consecutive displacement signals, and slope and speed. Therefore, the spatial signals of height and movement parameters of paths are effective on heart rate signals, and slope affects the movement parameter of speed. As shown in Fig. 4, the average heart rate signal and the height of all individuals are depicted for the round trip path. The linear regression of both signals is also given. For path 1, the overall trend of both signals is increasing. In other words, the heart rate increases as the height increases. It is worth noting that there is an increasing heart rate pattern for all individuals in path 1.



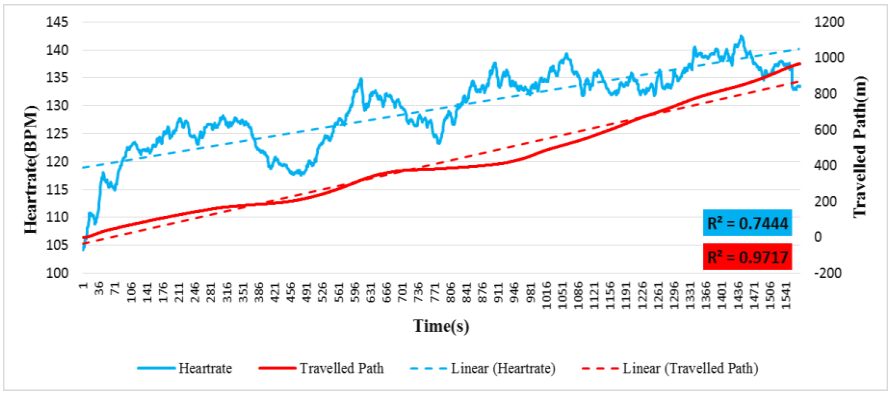

(a)



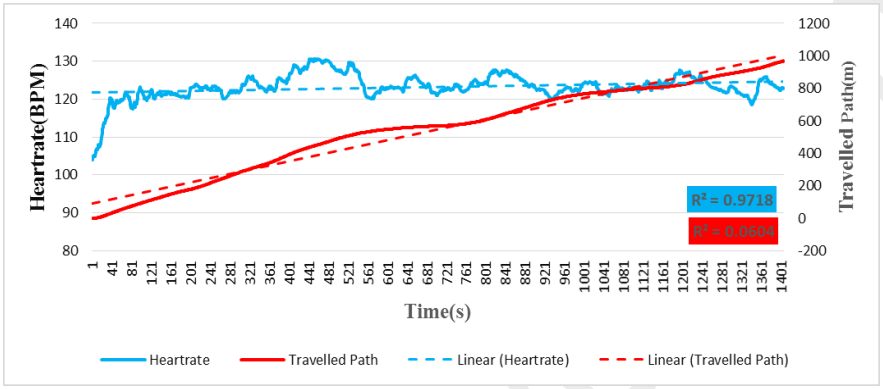

(b)

**Fig. 4** Relation between heart rate and height: (a) path 1, (b) path 2

As shown in Fig. 5, the average heart rate signals and paths of all individuals along with the linear regression of both signals are depicted for paths, respectively. As observed, the overall trend of both signals is increasing. This means that the heart rate increases as the traveled path increases. On path 2, there is an increased heart rate pattern for all individuals. However, an increase in heart rate for path 2 was lower than the increase in heart rate for path 1 because the height of path 2 is descending as observed in Fig. 4b.



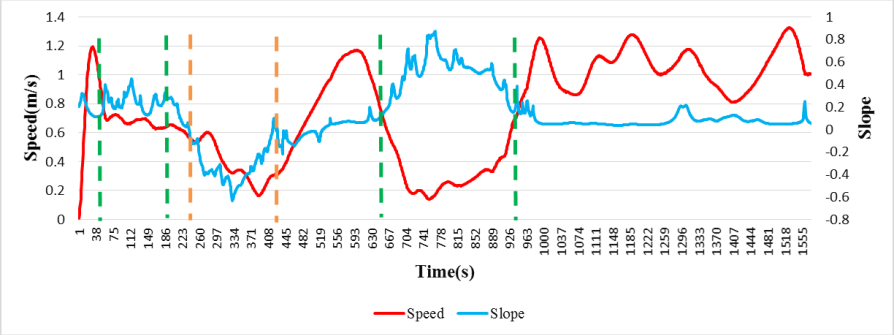

(a)



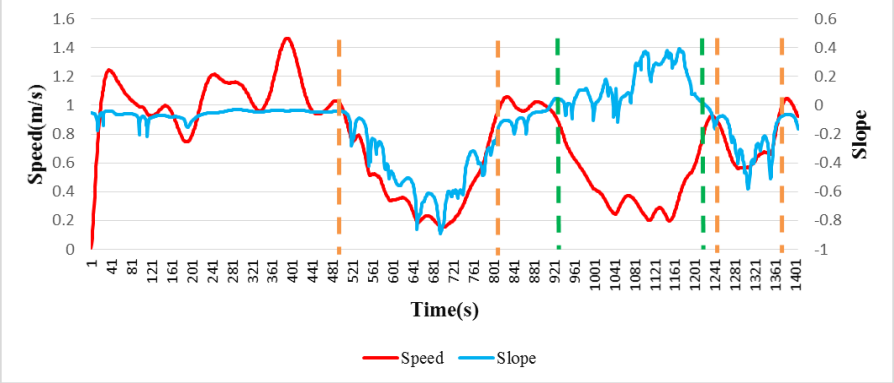

(b)

**Fig. 5** Relation between heart rate and traveled path: (a) path 1, (b) path 2

In addition, the average movement signal of speed for all the individuals and the slope for paths 1 and 2 are shown in Fig. 6. As shown, slope and speed generally have a level of relationship with each other. The part of the graph between the two green hachure lines indicates the uphill. Further, the part of the diagram between the two orange hachure lines indicates the downhill. It is worth noting that speed in the uphill and downhill decreases and the speed starts to increase when the slope of path decreases as long as it is zero. This trend exists for both paths. Interestingly, this is the pattern for all individuals.




(a)




(b)

**Fig. 6** Relation between slope and speed: (a) path 1, (b) path 2

### *4.2.2 Evaluating the performance of the proposed framework*

To evaluate the proposed framework, the obtained results from the similarity measurement function are compared to each other before and after reducing the number of signals. In this study, the results for path 1 indicated that the average difference between the obtained results before and after the dimension reduction is about 0.029 by choosing a mutual information threshold of 0.2. The same number for path 2 is 0.023.

In the mutual information method, removing dimensions is dependent on choosing a threshold. The threshold selection method is empirical. To select this value, the research approach is implemented by considering a certain threshold and the variation in the similarity value is evaluated. The similarity obtained from the process of similarity measurement before and after dimension reduction becomes less when a higher threshold is selected. For example, the average number of similarity variations is more than 0.3 by choosing a threshold less than 0.2. Furthermore, the average variation of similarity values is greater than 0.0005 by choosing a threshold less than 0.9. Regarding the variation in the similarity values before and after using the mutual information with different threshold limits, the value of 0.2 was selected as the threshold. This value can be set more than 0.2, in which case the variation in

similarity value is less than 0.03. However, it is possible to effectively remove a higher number of dimensions by choosing a threshold of 0.2 while achieving acceptable variations in similarity values.

The amount of mutual information of dimensions for path 1 and 2 is shown in Table 1 and 2. For example, the mutual information of heart rate and height dimensions of path 1 was calculated to be 0.231. By choosing a threshold of 0.2 for path 1, the *speed* dimension among the dimensions of consecutive displacement, slope, and speed, and the *heart rate* dimension among the dimensions of path, height, and heart rate enter the measurement process. The height, path, slope, and consecutive displacement dimensions are removed from the entities' dataset. However, the height dimension for path 2 is not removed.

**Table 1** Mutual information for path 1

| | **Speed** | **Consecutive displacement** | **Slope** | **Heart rate** | **Height** | **Traveled path** |
|---|---|---|---|---|---|---|
| **Speed** | 1 | 0.950 | 0.254 | - | - | - |
| **Consecutive displacement** | - | 1 | 0.253 | - | - | - |
| **Slope** | - | - | 1 | - | - | - |
| **Heart rate** | - | - | - | 1 | 0.231 | 0.377 |
| **Height** | - | - | - | - | 1 | 0.467 |
| **Traveled path** | - | - | - | - | | 1 |

**Table 2** Mutual information for path 2

| | **Speed** | **Consecutive displacement** | **Slope** | **Heart rate** | **Height** | **Traveled path** |
|---|---|---|---|---|---|---|
| **Speed** | 1 | 0.948 | 0.243 | - | - | - |
| **Consecutive displacement** | - | 1 | 0.242 | - | - | - |
| **Slope** | - | - | 1 | - | - | - |
| **Heart rate** | - | - | - | 1 | 0.190 | 0.341 |
| **Height** | - | - | - | - | 1 | 0.465 |
| **Traveled path** | - | - | - | - | - | 1 |

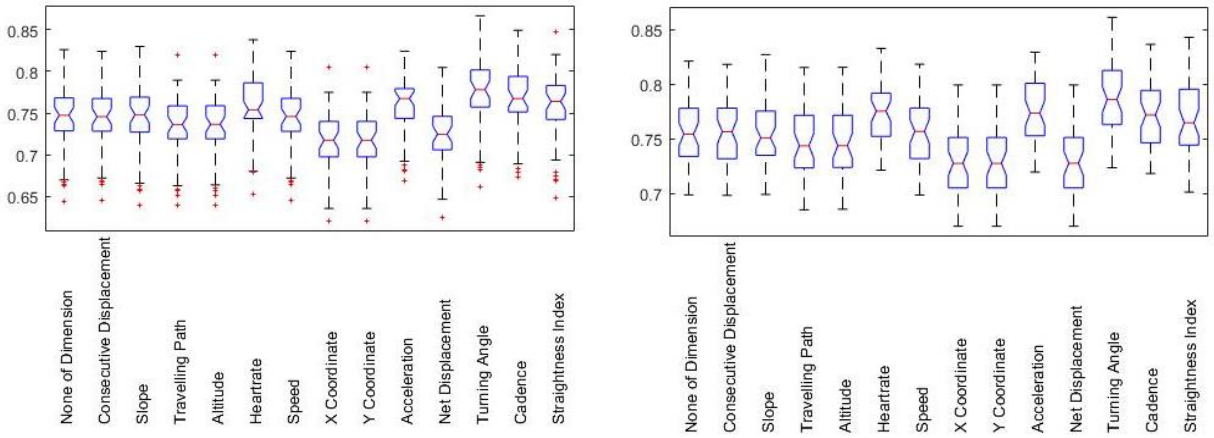
The difference in similarity due to the removal of the dependent dimensions is less than other dimensions because the remaining signals in the dataset can represent the dependent signals separately. For example, on path 1, the heart rate signal in the process of similarity measurement can represent the two dimensions of the height and path; and the speed signal can represent two dimensions of slope and consecutive displacement due to dependence. Furthermore, in path 2, the heart rate signal in the process of similarity measurement can represent path and the speed signal can represent two dimensions of slope and consecutive displacement due to dependence. Therefore, it is possible to remove dependent signals.

To provide a better understanding of the dependent signals, Table 3 presents the results of the similarity difference before and after the reduction of each signal. In fact, a two-by-two average similarity difference between all individuals before and after the reduction of each signal is represented. In each path, the data for 25 people were collected. Therefore, there are 45 comparative modes for every two individuals in each path. In this context, in each path, 45 similarity differences between two individuals are obtained before and after reducing each signal. Then, the average of these 45 values is calculated and given in Table 3. As observed, the obtained changes resulting from removing dependent dimensions are much less than other dimensions because there is a dimension in the dataset which has the same effect on the similarity value as its dependent dimension after removing the dependent dimensions. However, in the case of removing other dimensions, more variations are available in the similarity value. For example, by removing the x independent spatial dimension, the similarity value for both paths varies -0.0275 while the similarity value varies -0.0075 by removing the dependent dimension for path 1 and -0.0091 for path 2. The variations after removing the dependent dimension of paths are far less than removing the x independent spatial dimension.

**Table 3** Results of the similarity differences before and after reducing each of the signals

| | | Spatial (x) | Spatial (y) | Spatial (z) | Speed | Acceleration | Direct displacement | Consecutive displacement | Return angle | Heart rate | Traveled path | Movement rhythm | Direct value | Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Similarity of signals** | **Path 1** | -0.0275 | -0.0275 | -0.0073 | 0.0015 | 0.0166 | -0.0198 | 0.0014 | 0.0297 | 0.0149 | -0.0075 | 0.0253 | 0.0141 | 0.0029 |
| | **Path 2** | -0.0265 | -0.0265 | -0.0088 | 0.0015 | 0.0202 | -0.0265 | 0.0015 | 0.0320 | 0.0179 | -0.0091 | 0.0164 | 0.0142 | 0.0030 |

The boxplots of a two-by-two similarity measurement of all individuals before and after the reduction of each signal for both paths are presented in Fig. 7. The first boxplot is named "None of Dimension" because none of the dimensions is removed and it is depicted for the two-by-two similarity of all individuals without removing any of the dimensions in both shapes. However, one of the dimensions is removed and is named after the deleted dimension in the next boxplots. The boxplots related to the removal of the dependent dimension of consecutive displacement, slope, traveled path, and height for path 1 had very little changes compared to the first boxplot, which indicates little variation of similarity values. The boxplots related to the removal of independent dimensions for path 1 had more changes compared to the first boxplot, which indicates more variations of similarity values. For example, as shown in Fig. 7, the boxplot associated with the removal of the dependent consecutive displacement shows very slight changes compared with the first boxplot. However, the boxplot related to removing the independent dimension of acceleration indicates much more displacement and changes in comparison with the first boxplot. The above-mentioned cases are also true with a slight difference for path 2. In other words, the amount of variations in the boxplots for path 2 and path 1 is slightly different because the mutual information is different for these paths, leading to the variations in the similarity values for each path.




(a) (b)

**Fig. 7** Results of similarity measure function after and before each signal reduction for (a) path 1, (b) path 2

## 5 Discussion

As it was already mentioned, an individual's movement speed is affected by the path slope. This is clearly observable on round trips, especially on the uphill and downhill. Further, for path 1 the heart rate is partly influenced by the traveled path and height. This connection is weaker for path 2 because the height of this path is descending, which can have less fatigue for pedestrians.

In addition, there can be dependent signals in the collected dataset for each moving entity. The inclusion of these dependent signals in the process of similarity measurement should be avoided because they are affected by other spatial movement signals. These dependent signals can be identified using the unsupervised feature selection technique of mutual information. Table 1 and 2 indicate the values related to the mutual information of dependent signals. The remaining signals in the dataset can represent the dependent signals in the similarity measurement process. The similarity values before and after the reduction of dependent dimensions have little variations depending on the selection of the threshold. The average similarity variations from the comparison of both moving entities are presented in Table 3. Removing dependent dimensions of consecutive displacement, path, height, and slope caused less variation in the value of final similarity than in other dimensions. This is because the heart rate dimension represents the two dimensions of path and height and the speed dimension represents the two consecutive displacement and slope dimensions. Two dimensions of heart rate and speed have approximately the same effect as the final similarity due to their dependence on the mentioned dimensions.

The results of the similarity measurement function before and after the reduction of each of the signals for both paths indicate fewer changes after removing the dependent signals of consecutive displacement, slope, and path. Therefore, it is possible to consider all context dimensions in the similarity measurement process and avoid the inclusion of dependent dimensions which contain information affected by other dimensions effectively.

Choosing the appropriate threshold for the intended application is considered as an important issue in implementation of the research approach. In applications where a similarity variation of 0.03 is considered high, it is recommended to use higher thresholds. Furthermore, setting the threshold can be challenging in such a way that the height signal for path 2 cannot be removed by selecting a threshold of 0.2. However, the height and heart rate signal for path 2 have mutual information of 0.19.

## 6 Conclusion

Dependent movement, context dimensions and their relations had not been precisely considered in most movement studies. Due to significance of these issues in similarity measurement, attempts were made to identify and remove dependent dimensions by using one of the unsupervised feature selection techniques. Therefore, the present study introduced a novel context-aware framework for similarity measurement of the trajectories by only taking independent movement data and context information into account. The study aimed to specifically investigate the relationship between movement and context parameters in health indexes to reflect the behavior of affecting factors on health indicators. To this aim, a new approach was introduced for similarity measurement of trajectories, which could prevent the simultaneous effect of dependent dimensions on the dataset. In addition, the methods based on shapes and behaviors of signals, which have not been spatially used so far, were used in the calculation of the similarity of two signals. The implementation of these methods is relatively simpler than the ones introduced in the literature and the calculation of similarity is less time consuming. Finally, the dependent dimensions were removed using the unsupervised feature selection technique of mutual information. These issues will be very helpful in applications such as health recommender systems in which quick decision-making is a basic requirement.

According to the proposed framework results, it is possible to identify and remove dependent dimensions using the unsupervised feature selection method in dimension reduction techniques, so that the results of the similarity measurement process can have very little

variations. This framework can be evaluated in other studies with spatial trajectories and signals of vehicles or animals along with different context information. Furthermore, the findings of this research can be used in classifying and pattern discovering, and evaluating the performance of the research approach in different areas.

## Acknowledgements

The third author acknowledges the funding received from the Wittgenstein Prize, Austrian Science Fund (FWF), grant no. Z 342-N31.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Reference

Alvares, L. O., Bogorny, V., Kuijpers, B., De Macedo, J. a. F., Moelans, B.and Vaisman, A. A model for enriching trajectories with semantic geographical information. Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, 2007. ACM, 22.

Amouzandeh, K., Goudarzi, S.and Karimipour, F. Contextual Analysis of Spatio-Temporal Walking Observations. 2018 Cham. Springer International Publishing, 461-471.

Basiri, A., Amirian, P., Winstanley, A.and Moore, T. (2018). Making tourist guidance systems more intelligent, adaptive and personalised using crowd sourced movement data. *Journal of Ambient Intelligence and Humanized Computing,* 9**,** 413-427.

Brum-Bastos, V. S., Long, J. A.and Demšar, U. (2018). Weather effects on human mobility: a study using multi-channel sequence analysis. *Computers, Environment and Urban Systems,* 71**,** 131-152.

Buchin, M., Dodge, S.and Speckmann, B. (2014). Similarity of trajectories taking into account geographic context. *Journal of Spatial Information Science,* 2014**,** 101-124.

Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A.and Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research,* 10**,** 747-776.

Coelho, F., Braga, A. P.and Verleysen, M. (2016). A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems. *International Journal of Computational Intelligence Systems,* 9**,** 726-733.

Cover, T. M.and Thomas, J. A. (2012.) *Elements of information theory*, John Wiley & Sons.

Dai, J., Ding, Z.-M.and Xu, J.-J. (2016). Context-Based Moving Object Trajectory Uncertainty Reduction and Ranking in Road Network. *Journal of Computer Science and Technology,* 31**,** 167-184.

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

Dodge, S. (2019). A data science framework for movement. *Geographical Analysis,* 0**,** 1-21.

Dodge, S., Laube, P.and Weibel, R. (2012). Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science,* 26**,** 1563-1588.

Dodge, S., Weibel, R.and 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**,** 419-434.

Duch, W. (2000). Similarity-based methods: a general framework for classification, approximation and association.

Furtado, A. S., Kopanaki, D., Alvares, L. O.and Bogorny, V. (2016). Multidimensional similarity measuring for semantic trajectories. *Transactions in GIS,* 20**,** 280-298.

Giannotti, F.and Pedreschi, D. 2008. Mobility, data mining and privacy: A vision of convergence. *Mobility, data mining and privacy*. Springer.

Guessoum, D., Miraoui, M.and Tadj, C. (2016). Contextual location prediction using spatio-temporal clustering. *International Journal of Pervasive Computing and Communications,* 12**,** 290-309.

Hasanlou, M., Samadzadegan, F.and Homayouni, S. (2015). SVM-based hyperspectral image classification using intrinsic dimension. *Arabian Journal of Geosciences,* 8**,** 477-487.

Hosseini, R. S., Homayouni, S.and Safari, R. (2012). Modified algorithm based on support vector machines for classification of hyperspectral images in a similarity space. *Journal of Applied Remote Sensing,* 6**,** 063550.

Houari, R., Bounceur, A., Kechadi, M.-T., Tari, A.-K.and Euler, R. (2016). Dimensionality reduction in data mining: A Copula approach. *Expert Systems with Applications,* 64**,** 247-260.

Jun, J., Guensler, R.and Ogle, J. H. (2006). Smoothing methods to minimize impact of global positioning system random error on travel distance, speed, and acceleration profile estimates. *Transportation Research Record,* 1972**,** 141-150.

Kaffash-Charandabi, N., Alesheikh, A. A.and Sharif, M. (2019). A ubiquitous asthma monitoring framework based on ambient air pollutants and individuals' contexts. *Environmental Science and Pollution Research***,** 7525-7539.

Kays, R., Crofoot, M. C., Jetz, W.and Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science,* 348**,** aaa2478.

Krızek, P. (2008). *Feature selection: stability, algorithms, and evaluation.* Ph. d. thesis, Czech Technical University in Prague.

Laube, P. (2014.) *Computational movement analysis*, Springer.

Liu, S., Bruzzone, L., Bovolo, F.and Du, P. (2015). Hierarchical unsupervised change detection in multitemporal hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing,* 53**,** 244-260.

Nalmpantis, C.and Vrakas, D. Signal2Vec: Time Series Embedding Representation. International Conference on Engineering Applications of Neural Networks, 2019. Springer, 80-90.

Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D.and Smouse, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences,* 105**,** 19052-19059.

Ranacher, P.and Tzavella, K. (2014). How to compare movement? A review of physical movement similarity measures in geographic information science and beyond. *Cartography and Geographic Information Science,* 41**,** 286-307.

Ren, J., Zabalza, J., Marshall, S.and Zheng, J. (2014). Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. *IEEE Signal Processing Magazine,* 31**,** 149-154.

Ren, W., Song, J., Zhang, X.and Cai, X. (2016). Registration of multitemporal low-resolution synthetic aperture radar images based on a new similarity measure. *Journal of Applied Remote Sensing,* 10**,** 015001.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PloS one,* 9**,** e87357.

Seydi, S. T.and Hasanlou, M. (2017). A new land-cover match-based change detection for hyperspectral imagery. *European Journal of Remote Sensing,* 50**,** 517-533.

Sharif, M.and Alesheikh, A. A. (2017). Context-awareness in Similarity Measures and Pattern Discoveries of Trajectories: A Context-based Dynamic Time Warping Method. *GIScience & Remote Sensing,* 54**,** 426-452.

Sharif, M.and Alesheikh, A. A. (2018). Context-aware movement analytics: implications, taxonomy, and design framework. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 8**,** e1233.

Sharif, M., Alesheikh, A. A.and Tashayo, B. Similarity measure of trajectories using contextual information and fuzzy approach. Adjunct Proceedings of the 14th International Conference on Location Based Services, 2018. ETH Zurich, 75-80.

Sharif, M., Alesheikh, A. A.and Tashayo, B. (2019). CaFIRST: A context-aware hybrid fuzzy inference system for the similarity measure of multivariate trajectories. *Journal of Intelligent & Fuzzy Systems,* 36**,** 5383-5395.

Sharif, M.and Sadeghi-Niaraki, A. (2017). Ubiquitous sensor network simulation and emulation environments: A survey. *Journal of Network and Computer Applications,* 93**,** 150-181.

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U.and Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science,* 30**,** 881-906.

Stone, J. V. (2015.) *Information theory: a tutorial introduction*, Sebtel Press.

Wang, B., Wang, X.and Chen, Z. (2012). Spatial entropy based mutual information in hyperspectral band selection for supervised classification. *International Journal of Numerical Analysis & Modeling,* 9.

Wang, J. (2012). On the relationship between Pearson correlation coefficient and Kendall's tau under bivariate homogeneous shock model. *ISRN Probability and Statistics,* 2012.

Wang, Y., Deng, J., Gao, J.and Zhang, P. (2017). A hybrid user similarity model for collaborative filtering. *Information Sciences,* 418-419**,** 102-118.

Xia, Y., Wang, G.-Y., Zhang, X., Kim, G.-B.and Bae, H.-Y. Research of spatio-temporal similarity measure on network constrained trajectory data. International Conference on Rough Sets and Knowledge Technology, 2010. Springer, 491-498.

Yan, Z.and Spaccapietra, S. Towards semantic trajectory data analysis: A conceptual and computational approach. VLDB PhD Workshop, 2009. Citeseer.

Yuan, Y.and Raubal, M. (2014). Measuring similarity of mobile phone user trajectories–a Spatio-temporal Edit Distance method. *International Journal of Geographical Information Science,* 28**,** 496-520.

Zhang, J., Cao, Y., Zhuo, L., Wang, C.and Zhou, Q. (2015). Improved band similarity-based hyperspectral imagery band selection for target detection. *Journal of Applied Remote Sensing,* 9**,** 095091.

Zhao, W.and Du, S. (2016). Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Transactions on Geoscience and Remote Sensing,* 54**,** 4544-4554.

Zheng, Y.and Zhou, X. (2011.) *Computing with spatial trajectories*, Springer Science & Business Media.