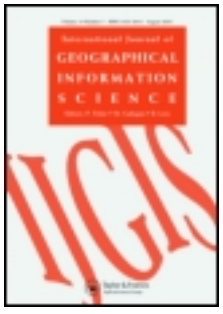

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Measuring similarity of mobile phone user trajectories - a Spatio-temporal Edit Distance method

Yihong Yuan[ab] & Martin Raubal[a]

[a] Institute of Cartography and Geoinformation, ETH Zurich, 8093 Zurich, Switzerland

[b] Department of Geography, University of California, Santa Barbara, CA 93106, USA

Published online: 17 Dec 2013.

To cite this article: Yihong Yuan & Martin Raubal , International Journal of Geographical Information Science (2013): Measuring similarity of mobile phone user trajectories - a Spatio-temporal Edit Distance method, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2013.854369

To link to this article: http://dx.doi.org/10.1080/13658816.2013.854369

International Journal of Geographical Information Science, 2013
http://dx.doi.org/10.1080/13658816.2013.854369

Taylor & Francis
Taylor & Francis Group

Measuring similarity of mobile phone user trajectories – a Spatio-temporal Edit Distance method

Yihong Yuan[a,b]* and Martin Raubal[a]

[a]*Institute of Cartography and Geoinformation, ETH Zurich, 8093 Zurich, Switzerland;* [b]*Department of Geography, University of California, Santa Barbara, CA 93106, USA*

(*Received 31 March 2013; final version received 7 October 2013*)

The rapid development of information and communication technologies (ICTs) has provided rich data sources for analyzing, modeling, and interpreting human mobility patterns. This paper contributes to this research area by developing the Spatio-temporal Edit Distance measure, an extended algorithm to determine the similarity between user trajectories based on call detailed records (CDRs). We improve the traditional Edit Distance algorithm by incorporating both spatial and temporal information into the cost functions. The extended algorithm can preserve both space and time information from string-formatted CDR data. The novel method is applied to a large data set from Northeast China in order to test its effectiveness. Three types of analyses are presented for scenarios with and without the effect of time: (1) Edit Distance with spatial information; (2) Edit Distance with time as a factor in the cost function; and (3) Edit Distance with time as a constraint in partitioning trajectories. The outcomes of this research contribute to both methodological and empirical perspectives. The extended algorithm performs well for measuring low-resolution tracking information in CDRs, as well as facilitating the interpretation of user mobility patterns in the age of instant access.

Keywords: human mobility patterns; trajectory similarity measure; mobile phone data sets; edit distance; time series

1. Introduction

Modeling human mobility patterns has become a crucial research topic in various fields such as Physics, Transportation, and Geographic Information Science (Kwan 2000, Du Mouza and Rigaux 2005, González *et al*. 2008). Much progress has been made regarding the theories, methodologies, and applications in this area (Kwan 2004, Miller and Han 2009, Song *et al*. 2010, Liu *et al*. 2012, Yuan and Raubal 2012b). Due to the natural variability of individual mobility and the uncertain data quality of recorded location information, analyzing human trajectories has been a challenging research question in spatio-temporal data mining and knowledge discovery. Researchers have focused on different aspects in this field, including *intra-trajectory* studies, i.e., understanding the internal regularity of human motions (González *et al*. 2008), and *inter-trajectory* studies, i.e., measuring trajectory similarity between individuals (Xia *et al*. 2011). The latter has drawn more and more attention due to the increasing interest in understanding the social interaction among demographic groups (Joh *et al*. 2002, Eagle *et al*. 2009). Measuring

*Corresponding author. Email: yyuan@ethz.ch

trajectory similarity can also support many real-world applications, such as traffic analysis or crime prediction (Zheng and Zhou 2011, Baron 2012).

In addition, the rapid development of information and communication technologies (ICTs) has introduced a wide range of novel spatio-temporal data sources (e.g., georeferenced mobile phone records) for researchers to explore the movement patterns of its carriers (Ahas *et al*. 2010). Although mobile phones are capable of recording location information through several ways such as assisted global positioning system (GPS), the collected data are normally generated as scattered points in call detailed records (CDRs) which include the IDs of connected cell towers for call activities. These data can be viewed as strings of cell IDs characterized by low accuracy and precision in both space and time dimensions. Although CDRs cannot represent the accurate locations of phone carriers, these data can be considered as an approximation/sampling of the real trajectories. Previous research has demonstrated the effectiveness of modeling user activity patterns based on CDR data (Ahas 2005, Eagle *et al*. 2009). However, there has not been sufficient research on how to investigate the similarity between user trajectories based on these scattered and cell-based sample points. As argued in Kang *et al*. (2009), a similarity measure for cellular space is different from one for Euclidean space because numerical information for cellular space is not necessarily continuous. Therefore most of the existing algorithms, such as the time Synchronized Euclidean Distance (SED) and the Hausdorff distance (Zheng and Zhou 2011), are not easily applicable. In the field of mobile radio technology, cells represent the smallest units of a given space. Although in practice the signal coverage of cells can overlap with each other, cellular space is usually simplified in previous research as units that can touch but not overlap with each other (Kolbe *et al.* 2008). A common conceptualization of these cells are Voronoi polygons (Sharifzadeh and Shahabi 2006), which define a way of dividing space into a number of regions based on a set of points (seeds), so that the corresponding region consists of all points closer to that seed than to any other. For mobile phone data, the locations of base towers are usually considered as seeds for generating a Voronoi diagram (Baert and Seme 2004, Stergiopoulos and Tzes 2009). In this research we focus on measuring trajectory similarity based on a Spatio-temporal Edit Distance algorithm, which was in its initial version proposed for string matching and correction in the 1970s (Wagner and Fischer 1974). This method belongs to the family of sequence alignment algorithms (Abbott 1995). It calculates the minimum number of operations required to switch one string to another; therefore it is highly suitable for matching series of cell IDs in CDRs. Another advantage of this method is that it can deal with sequences of different lengths (Wilkes 2008), which is a typical situation in CDRs. However, traditional Edit Distance deals with purely qualitative information such as alphabet letters. In order to preserve the spatial information in CDRs, we will modify the cost function of the algorithm to incorporate the spatial distribution of cell towers. The modified algorithm will be helpful for measuring low-resolution tracking information in CDRs, as well as facilitating the interpretation of user mobility patterns in the age of instant access. Moreover, since the temporal aspect is of major importance in human mobility, we will also present two exemplary analyses which explicitly incorporate the effect of time.

The remainder of this paper is organized as follows: Section 2 describes related work in the areas of human mobility, trajectory similarity measures, and the Edit Distance algorithm. In Section 3, we introduce the basic research design, including the description of the data set and the methodology. Section 4 presents the three variations of data analysis based on the Extended Edit Distance method. In Section 5, we discuss the results and further indications. Section 6 presents conclusions and directions for future research.

2. Background

2.1. Human mobility and trajectory similarity measures

Modeling and interpreting human trajectories has been a challenging research question due to the complex nature of human motions and the diverse formats of the recorded location information. Larsen *et al*. (2006) identified five types of mobility: (1) Physical travel of people (e.g., work, leisure, family life); (2) Physical travel of objects (e.g., products to customers); (3) Imagination travel (e.g., memories, books, movies); (4) Visual travel (e.g., Internet surfing on Google Earth); and (5) Communication travel (e.g., person-to-person messages via telephones, letters, emails, etc.). However, these five types of mobility are not independent. In this research, when referring to 'human mobility' we focus on characterizing *Physical travel of people* (*trajectories*) from records of *Communication travel* (*CDRs*).

Zheng and Zhou (2011) summarized existing research questions in trajectory analysis and divided them into three categories: trajectory preprocessing (prior to the database-level), trajectory indexing and retrieval (in databases), and advanced topics (above the database-level). One of the advanced topics for analyzing trajectories is determining their similarity/dissimilarity to each other. Researchers have investigated several methods to quantify how similar two trajectories behave in spatial and/or temporal dimensions, and these methods can be divided into two categories:

(1) Shape-based methods: This category eliminates the temporal aspect (i.e., speed) and only focuses on the geometric characteristics (i.e., shape) of trajectories. A trajectory can either be considered as a series of scattered visited points, or a polygonal line that may self-intersect and have duplicate vertices. Typical methodologies in this category include but not are limited to:

- ***Classical Euclidean Distance***: The most straightforward method which adds up the distance measures between each corresponding pair of points. However, this method requires the two compared sequences to have the same number of points, and is therefore not applicable to most real-world applications when two compared series have a varying number of points.
- ***Hausdorff Distance***: A shape comparison metric between two point sets, which determines the longest of all the distances from a point in one set to the closest point in the other set (Rucklidge 1997). This method is mostly applicable for measuring pair-wise distances without considering the sequences and directions of points.

(2) Time-based methods: Researchers have previously recognized the limitations of shape-based methods; therefore, another category of methods has been developed, which considers the role of temporal components in trajectories. In these methods the compared features are considered as multidimensional time series data, and can be processed by techniques extended from time series analysis and sequence comparison (Buchin *et al*. 2009), such as:

- ***Synchronized Euclidean Distance***. This algorithm uses the classic Euclidean Distance by measuring the distance between two points at identical time stamps (Potamias *et al*. 2006). However, this method does not perform well with distortions and replications in trajectories.

- ***Discrete Fréchet Distance* (*aka the Coupling Distance*)**. A measure for the similarity between two curves which considers the location and ordering of the points along the curves; however as demonstrated by previous studies, this method is very sensitive to outliers and displacements (Eiter and Mannila 1994).
- ***Dynamic Time Warping* (*DTW*)**. This method has been well developed in the field of speech recognition, signal processing, and related sequence measures in one-dimensional space (Senin 2008). Due to the fact that the one-dimensional calculation in this algorithm can be easily replaced by distance measures for spatial points, DTW can be extended to measure the trajectory similarity of human motion (Makinen 2001, Yuan and Raubal 2012a). This method is mostly applicable to deal with time series with potentially large distortions in the time dimension; however, it is not specifically suitable for string-formatted CDR data. The computing time load is also heavy compared to other algorithms.
- ***Longest Common Sub-Sequence* (*LCSS*)**. This method aims to find the longest common subsequence in a set of sequences. Its basic idea is to match the sequences allowing for the elimination of outliers (i.e., some elements remain unmatched), which can be considered as a special case of the Edit Distance method (Maier 1978). However, it is not very suitable for comparing human trajectories due to the fact that outliers in trajectories may also have a significant impact on the exploration of movement patterns.

Although the majority of existing similarity measures are based on Euclidean space, researchers have also explored the comparison of trajectories in Non-Euclidean space (i.e., network-based or cell-based space). For example, Won *et al*. (2009) proposed a new scheme for trajectory clustering in road network space which judges the degree of similarity by considering the total length of matched road segments. With the widespread usage of mobile location-aware devices, a large amount of trajectory data are captured every day and stored in CDRs. In this type of records, the trajectory of phone users can be considered as a sequence of visited cell tower IDs in cellular space; meanwhile each tower is georeferenced by geographic coordinates; therefore, traditional similarity measures in Euclidean space are not sufficient for this type of data (semi-qualitative, semi-quantitative). Related research can be found in Kang *et al*. (2009), where the authors conducted trajectory clustering in a cellular space. However, they only considered the sequence of cell IDs as a regular string without taking into account the spatial distribution of these cells. Shoval and Isaacson (2007) also demonstrated the potential of the Sequential Alignment Method in comparing trajectories, but their method has the same problem of not taking into account the real coordinates for each site. Other closely related research was conducted by Dodge *et al*. (2012), where trajectories are separated into segments with specific movement parameters (MPs) such as velocity. The authors used alphabetical letters to denote different MP classes, i.e., the original trajectories are converted to string sequences (i.e., *ABADCB*), then a modified version of the Edit Distance was applied to compute the similarity between two MP sequences. This method focuses on measuring the similarity based on selected MPs, therefore, it is most appropriate when the space-time geometry of the movement is not the major focus of the analysis (Long and Nelson 2013). As mentioned in Section 1, CDR data are often provided with low data quality, and it is difficult to extract reliable MP classes. Therefore, this method is not directly applicable to the data used here.

In addition to similarity measures, it is often essential to classify spatial objects into subgroups in practice, so that objects within the same group are more similar to each other than those in different groups (Miller 2009). These techniques have also been applied in the field of trajectory analysis (Lee *et al*. 2007a), which concentrates on classification and clustering of multiple trajectories based on their shapes and other features. Researchers have developed various spatio-temporal clustering techniques for grouping observations that show similar behavior in both spatial and temporal dimensions, from basic clustering techniques such as *k*-means clustering and hierarchical clustering to more advanced techniques such as Hidden Markov Model (HMM) and Principle Component Analysis (PCA) (Lee *et al*. 2001). As argued by Salvador and Chan (2004), one essential problem in clustering analysis is to determine the number of clusters. Several methodologies have been proposed to tackle this issue, including the Elbow (distinctive break) method, Gap statistics, etc. (Tibshirani *et al*. 2001), among which the Elbow method has been widely adopted due to its simplicity to illustrate and implement (Foss and Zaïane 2002, Salvador and Chan 2004).

In this research, to take CDR data analysis one step further, we extend the traditional Edit Distance algorithm by incorporating the spatial distribution of cell towers, and then apply the newly developed Spatio-temporal Edit Distance to compare trajectories extracted from CDRs and conduct clustering analysis. As mentioned in Section 1, Edit Distance is not a typical choice for measuring the similarity between human trajectories, since it was mainly used for string matching; however, it is highly applicable to CDR data due to the fact that CDRs are often string-formatted. Section 2.2 provides an overview of the Edit Distance algorithm, and the detailed methodology will be elaborated in Section 3.

2.2. *Edit Distance algorithm and its applications*

The methodology in this research is based on the Edit Distance algorithm proposed by Wagner and Fischer (1974), which measures the distance between two strings by computing the number of edit operations when transforming one string to another. The pseudocode is shown as follows:

Given two strings S (s_1, …, s_i) and T (t_1, …, t_j), in the optimal case, to transform S to T there are three solutions:

- s_i is deleted and the rest s_1, …, $s_{i\text{-}1}$ is transformed to t_1, …, t_j,
- s_1, …, s_i is transformed into t_1, …, $t_{j\text{-}1}$ and we insert t_j at the end
- s_i is changed into t_j and the rest s_1, …, $s_{i\text{-}1}$ is transformed to t_1, …, $t_{j\text{-}1}$.

Thus the recursive algorithm can be defined as

$$\begin{aligned}\text{EditDistance}[i,j] = \min(&\text{EditDistance}[i-1,j] + \text{Cost}[\text{delete}\,(s_i)],\ \text{EditDistance}\\ &[i,j-1] + \text{Cost}\big[\text{insert}\,(t_j)\big],\ \text{EditDistance}[i-1,j-1] + \text{Cost}\big[\text{replace}\big(s_i,t_j\big)\big]\big).\end{aligned}$$

In string matching, the cost of each operation is usually set as constant 1, whereas in real-world applications it is defined based on practical needs. For instance, linear gap-costs are sometimes used where a run of insertions (or deletions) of length 'x' has a cost of '$mx + n$' ('m' and 'n' are constants), indicating that if n is larger than 0, this penalizes short runs of insertions and deletions (Powell *et al*. 2000). Researchers have also proposed variations of this method, such as the Edit Distance on real sequence (EDR), the Edit Distance with real penalty (EDRP) (Chen *et al*. 2003), and the Extended Edit Distance (EED) (Marwan *et al*. 2008).

The Edit Distance algorithm has also been applied in other fields besides spelling correction. For example, Yang and Tiow (2007) employed this method for remote screen updates to measuring the differences between a 'picture' of what the screen currently is and another picture of what it should become. In Molecular Biology, Edit Distance is utilized to test how similar two DNA sequences are (Smith and Waterman 1981). It has also been applied to plagiarism detection (Zini *et al*. 2006).

In this research we will extend the classic Edit Distance method. The traditional algorithm calculates the minimum number of operations when converting one sequence to another. We take this method one step further by incorporating the locations of the cell towers. This eliminates the disadvantage that every operation has equal influence (the same cost function). Moreover, previous research has considered time as a third dimension when constructing the space-time paths of individuals (Miller 2005, Kwan 2006). Classic Edit Distance considers the order of how the points are aligned, but the effect of time is not explicitly represented. Here we also investigate the effect of time in the modified cost function. Although Lee *et al*. (2007b) used the same term 'Spatio-temporal Edit Distance (STED)' to compare moving objects in video surveillance, our approach adopts a different perspective in modifying the cost function. Compared to their vector-based approach, our algorithm focuses on the impact of each operation on the overall spatial distribution of trajectories. Based on this extended method, it is also feasible to explicitly adjust the weights of spatial and temporal components in the cost function (Section 4.2.1). Besides CDR data, the proposed Spatio-temporal Edit Distance can similarly be applied to other data sources such as Bluetooth tracking data and location records restrained by road networks or landmarks.

3. Research design

3.1. Data set

For this research we utilize a data set from Northeast China, which covers over 1.7 million people and includes CDRs for a time span of 9 days (5 weekdays, 4 weekend days) in City *A*.[1] It includes the time, duration, and the location of the corresponding cell tower for each mobile phone connection. The data set only covers voice calls (not including other connections such as text messages or Internet connection). For each user, the CDR data record the location of the nearest mobile phone tower when the user initiates or receives a phone call. Based on the spatial density of cell towers, positional data accuracy is about 300–500 m. Table 1 provides a sample record. The phone numbers, cell IDs, longitudes, and latitudes are not shown for reasons of privacy. Note that the location records in the data set cannot represent the exact trajectory of each user due to both resolution (only recorded when a call connection has been established) and accuracy (only the nearest tower locations are recorded) issues. However, as argued in González *et al*. (2008), the

Table 1. Sample record from the data set.

Phone #	13601******
Cell ID	01**
Cell Longitude	126.*****
Cell Latitude	45.*****
Time	16:10:31
Duration	11 minutes

mobility of phone users indicates a high level of regularity based on a time cycle of 10 days. It is highly probable for an individual to return to the location where he/she was first observed within the following 240 hours. Therefore, based on a summary of 9 days' records, the data in this research can be utilized to depict the general characteristics of user trajectory patterns (Yuan *et al.* 2012).

3.2. Methodology

Given two trajectories $R_1[p_{11}(x_{11}, y_{11}, t_{11}), \ldots, p_{1n}(x_{1n}, y_{1n}, t_{1n})]$ and $R_2[p_{21}(x_{21}, y_{21}, t_{21}), \ldots, p_{2m}(x_{2m}, y_{2m}, t_{2m})]$ (where p_{ij} represent space-time points, and each x_{ij}, y_{ij}, and t_{ij} represents the longitude and latitude of the connected tower, and the time of call activity), it is highly possible that users can make several phone calls at the same location within a short time interval; therefore, it is necessary to clean up the data by removing the redundant points before the analysis. In CDRs, it is highly possible that the temporal resolution of data is unevenly distributed. One user may attempt to establish a large number of call connections in a very short time span (i.e., due to weak signal or bad connection quality). For example, for a salesman, it is possible to make 10 phone calls within 30 minutes in the office; however, it is not sensible to keep all the data points and assume that the salesman visits the office 10 times within 30 minutes; hence eliminating these points helps us to reduce calculation bias and the time load of the algorithm. For instance, given three Users *A*, *B*, and *C*:

- *A* made 1 phone call at location *X*, 10 phone calls at location *Y* (within a very short time span, e.g., 30 minutes), and 1 phone call at location *Z*.
- *B* made 1 phone call at location *X*, 1 phone call at location *Y*, and 1 phone call at location *Z*.
- *C* made 10 phone calls at location *Y* (within a short time span, e.g., 30 minutes)

Based on the provided information, it is reasonable to assume that *B*'s trajectory is more similar to *A* (compared to User *C*) (i.e., distance (*A,B*) < distance (*A,C*)), since the 10 phone calls at location *Y* within 30 minutes can be considered as 'a single visit'. However, if we keep all 10 phone calls at location *Y* as 10 points, the sequences for *A*, *B*, and *C* are as follows:

$$A: XYYYYYYYYYYZ;\ B: XYZ;\ C: YYYYYYYYYY$$

Since the calculation of Edit Distance is highly dependent on the number of operations, the result will indicate that distance (*A,B*) > distance (*A,C*), which is contradictory to our common sense. Hence, it is necessary for us to eliminate redundant points within a short time span. Here the redundant points are defined as follows (Figure 1):

For any two consecutive points p_i and p_{i+1} in a given trajectory, if p_i and p_{i+1} are located within the cell of the same mobile phone tower, and the time difference $t_{i+1} - t_i < \Delta T$ (ΔT is a threshold value, in this paper predefined as 0.5 hour), p_{i+1} is defined as a redundant point and removed. Meanwhile t_i is updated as the average of t_i and t_{i+1} (for instance, in Figure 1, t_4 is deleted and t_3 is updated as 10:05).

As discussed in Section 2.1, in CDRs, the trajectory of a phone user can be represented by a sequence of cell IDs, for example, [Cell5, Cell6, Cell5, Cell4, Cell3, Cell5]. The distance between two trajectories can be measured by the cost of operations required to transform one sequence to the other. In practice, cost functions are often defined in a manual fashion, from

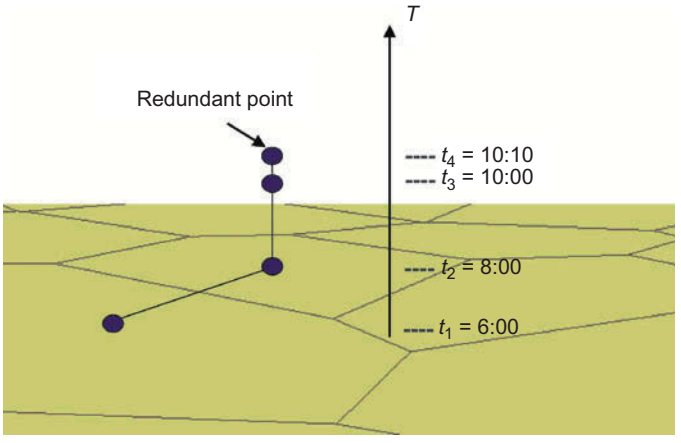

Figure 1. Removing redundant points in CDR-based trajectories.

commonly used constant values (Zini *et al*. 2006) to customized weighted functions (Powell *et al*. 2000). In this research, the tracking information in CDRs is not purely qualitative, indicating that each operation should be assigned a different cost value based on the locations of the deleted/inserted/replaced points. From a geometric perspective, trajectories can be considered as a set of finite points, where the centroid of a certain trajectory is calculated as the average location of these points. Since a centroid minimizes the sum of squared Euclidean distances between itself and each point in the set, it can be considered a 'balance point' for the whole trajectory (Johnson 2007), which is often used as a reference or benchmark point in spatial point pattern analysis, such as the Centroid Distance Function (Yang et al. 2008, 2012). Therefore, deleting a point that is faraway from the centroid leads to a higher impact on the spatial distribution of the original trajectories than deleting a nearby point. The main improvement of the extended algorithm is to assign the operation cost based on the impact of each operation by measuring the centroid displacement after each operation. The cost functions are defined as:

$$\text{Cost}[\text{Delete}\,(p_{1i})] = \sqrt{(1-c)\left\{\left[\left(\frac{\sum_{k=1}^{n} x_{1k}}{n}\right)-\left(\frac{\sum_{k=1,k\neq i}^{n} x_{1k}}{n-1}\right)\right]^2+\left[\left(\frac{\sum_{k=1}^{n} y_{1k}}{n}\right)-\left(\frac{\sum_{k=1,k\neq i}^{n} y_{1k}}{n-1}\right)\right]^2\right\}+c\left[\left(\frac{\sum_{k=1}^{n} t_{1k}}{n}\right)-\left(\frac{\sum_{k=1,k\neq i}^{n} t_{1k}}{n-1}\right)\right]^2} \quad (1)$$

(Displacement of the centroid in the trajectory R_1 after removing p_{1i});

$$\text{Cost}\,[\text{Insert}\,(p_{2j})] = \sqrt{(1-c)\left\{\left[\left(\frac{\sum_{k=1}^{n} x_{1k}}{n}\right)-\left(\frac{\sum_{k=1}^{n} x_{1k}+x_{2j}}{n+1}\right)\right]^2+\left[\left(\frac{\sum_{k=1}^{n} y_{1k}}{n}\right)-\left(\frac{\sum_{k=1}^{n} y_{1k}+y_{2j}}{n+1}\right)\right]^2\right\}+c\left[\left(\frac{\sum_{k=1}^{n} t_{1k}}{n}\right)-\left(\frac{\sum_{k=1}^{n} t_{1k}+t_{2j}}{n+1}\right)\right]^2} \quad (2)$$

(Displacement of the centroid in trajectory R_1 after inserting p_{2j});

$$\text{Cost}\,[\text{Replace}\,(p_{1i}, p_{2j})] = \sqrt{(1-c)\left\{\left[\left(\frac{\sum_{k=1}^{n} x_{1k}}{n}\right)-\left(\frac{\sum_{k=1,k\neq i}^{n} x_{1k}+x_{2j}}{n}\right)\right]^2+\left(\frac{\sum_{k=1}^{n} y_{1k}}{n}\right)-\left(\frac{\sum_{k=1,k\neq i}^{n} y_{1k}+y_{2j}}{n}\right)\right]\right\}^2+c\left[\left(\frac{\sum_{k=1}^{n} t_{1k}}{n}\right)-\left(\frac{\sum_{k=1,k\neq i}^{n} t_{1k}+t_{2j}}{n}\right)\right]^2} \quad (3)$$

(Displacement of the centroid in trajectory R_1 after replacing p_{1i} by p_{2j}).

Here we introduce a constant $c \in [0, 1]$ to balance the influence between the spatial and temporal dimensions. For $c = 0$ the equations become purely spatial, whereas when $c = 1$ the cost functions only consider the time when the activities happened and ignore the impact of the spatial dimension. This can be utilized for exploring the similarity between phone call occurrence patterns, i.e., to answer questions such as 'how similar is User A to B when only considering their phone call occurrence?' In this research spatial effect is our major focus; therefore, we start the analysis in Section 4.1 with $c = 0$. In Section 4.2 we also discuss the time effect when $c > 0$.

Note that two corresponding operations may have different weights in the cost function depending on which one of the two trajectories is considered the target, i.e., 'replacing p_{1i} by p_{2j} in R_1' may have a different impact compared to 'replacing p_{2j} by p_{1i} in R_2'. This may result in a slight difference between EditDistance(R_1, R_2) and EditDistance(R_2, R_1). The rationality of asymmetric distances has been discussed in Cognitive Science (Tversky 1977) and Geographic Information Science (Janowicz et al. 2011), but in this research we focus on the physical aspects of objects (instead of the cognitive aspects). Hence, to preserve the symmetry of distance calculation we choose the average of the two values.

As indicated in the methodology, the proposed algorithm incorporates the displacement of the centroid in the cost function. Compared to traditional methods of calculating the bounding box of trajectories (Schneider 1999, Han et al. 2004), it also considers the frequency and order of how different points are visited. Moreover, it improves upon the classic Edit Distance in which the length of sequences has a large impact on the results by (1) eliminating redundant points and (2) incorporating the spatially enabled cost functions. In the extended algorithm, if a point is near the centroid (i.e., a low-influence point), an operation on this point will not result in a large difference even if it appears repeatedly in the trajectories.

Although the spatial attribute is of major importance in this research, we are also interested in the effect of time in the cost functions. In Section 4 we will present three analyses to demonstrate the advantages of the proposed algorithm: (1) a generic similarity measure (when setting $c = 0$), which characterizes the similarity between trajectories without explicitly specifying the effect of time; (2) time-enabled comparison, which considers the time dimension as a parameter in the cost function ($c > 0$); and (3) time-enabled comparison, which introduces time as a constraint in partitioning trajectories.

4. Analysis and results

4.1. Generic analysis

As indicated in Section 3, in the generic analysis we introduce various applications of the proposed Spatio-temporal Edit Distance focusing on the effect of 'space' in the cost functions, including a trajectory comparison analysis and a clustering analysis. The former

one demonstrates how to identify users with the most similar trajectory patterns, as well as how to detect outlier trajectories based on average distances. The latter one shows how sample trajectories are grouped into clusters with different distribution patterns in both activity space and movement directions.

4.1.1. *Trajectory comparison*

In a first step, we need to eliminate those users from the data set who have too few call records. After data preprocessing, we obtained 844,784 users for weekdays and 675,832 users for weekend days whose CDRs include more than 10 records. Figure 2 shows the result of a comparison between an example User *A* and his/her most similar users based on our algorithm. For comparison, longitude and latitude values in our sample are scaled to the range [0, 1]. Trajectories *B* and *C* are the most similar trajectories to *A* on weekdays (Figure 2a, scaled distance = 0.0017) and weekend days (Figure 2b, scaled distance = 0.0035). As can be seen, the algorithm mainly focuses on matching the activity area of two trajectories. Moreover, it also considers the order of how the points are visited in time.

As indicated in Figure 2, the proposed method is effective in identifying the similarity of trajectory patterns in terms of identifying the range of activity space and the order of visited points. The result of this analysis can be useful for phone companies to interpret user activities, as well as improving algorithms in social network applications, such as 'friend recommendation' based on movement patterns.

For a better overview of the general trend in user trajectories, we randomly selected a smaller sample set of 1000 users, and calculated the average distance between each user and all the other 999 users (which also forms a distance matrix). This average distance can be considered as an indicator of how 'different' a user behaves in terms of mobility pattern compared to the others. As shown in the weekday histogram (Figure 3a, mean value = 0.148), the distance follows a skewed normal distribution, and the positive tail indicates that there is a certain number of users with a large average distance (>0.5). Moreover, the mean distance on weekend days (Figure 3b, mean value = 0.124) is tested to be smaller than the one on weekdays based on a two sample *t*-test (significance level $P = 0.05$), indicating that the trajectory patterns on weekend days are less diverse than those on weekdays. This is inconsistent with common sense that weekend activities are more random; however it confirms the results of a previous study by Yuan and Raubal (2012a).

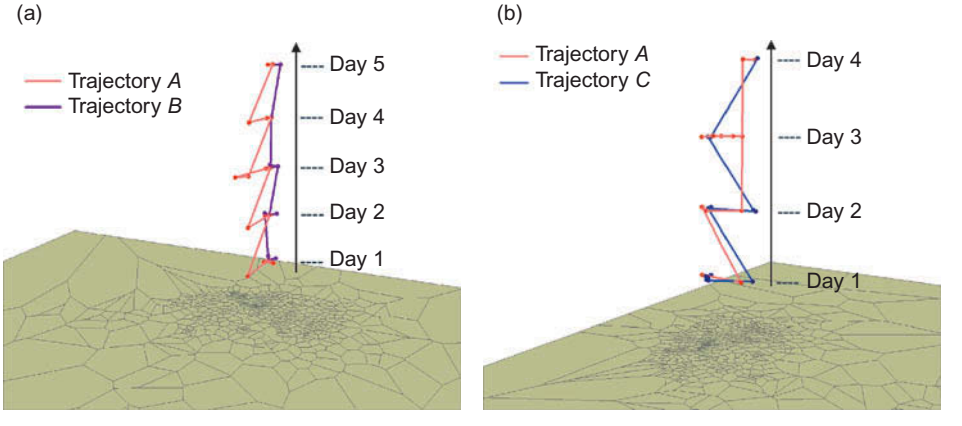

Figure 2. Example analysis: the most similar trajectories for (a) weekdays and (b) weekend days.

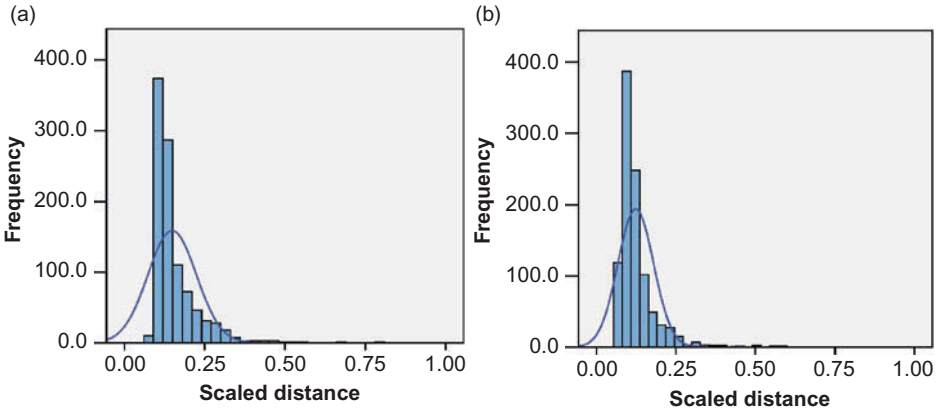

Figure 3. Histogram of average distance on (a) weekdays and (b) weekend days.

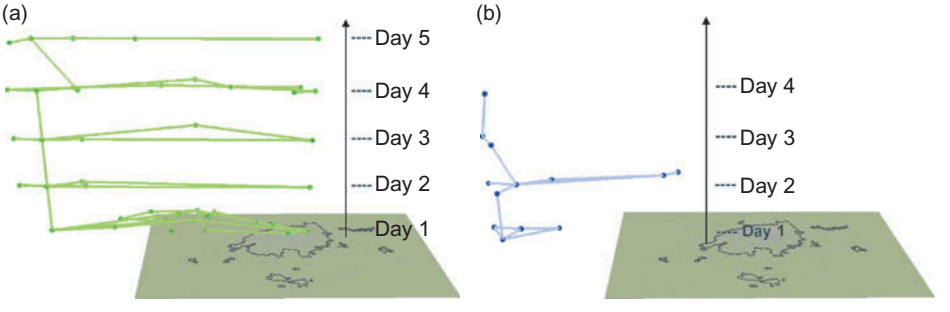

Figure 4. Users with the largest average distance on (a) weekdays (b) weekend days.

For illustration we also plot the spatial trajectories of the two users who have the largest average distance (one for weekdays and one for weekend days). As can be seen in Figure 4a, the user's activity area is quite large (some points are even located outside of the target area). Similarly, the pattern indicated in Figure 4b is distant from the city center with most activities outside of the study area (built-up areas in the city are delineated in Figure 4). Such analysis and visualization can be helpful for detecting abnormal patterns in mobile networks, as well as providing input for service providers and certain user groups who are interested in mobility pattern comparison.

4.1.2. Clustering analysis

The comparison analysis demonstrated that the proposed algorithm can be utilized to measure the similarity between any pair of user trajectories. For a group of users, it is useful to take a step further from similarity measure to clustering analysis. In the remainder of this section, in order to enhance the interpretation of the general pattern for the sample data set, we also conduct a hierarchical clustering analysis (here we use weekday patterns as an example).

To determine the number of clusters, a variation of the Elbow method is utilized and the number of clusters (on the x axis) is plotted against the merge distance (on the y axis).

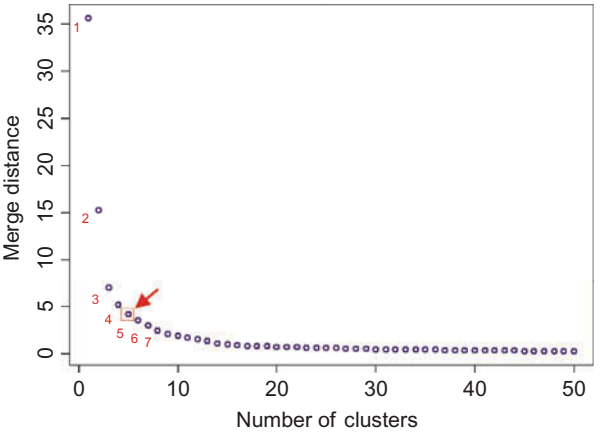

Figure 5. Determining the number of clusters (smoothed data).

The merge distance is defined as the differences between two merged clusters (in this case calculated as the increase in the 'error sum of squares' (ESS) after fusing two clusters into a single cluster based on the Ward's linkage method) (Ward 1963, Everitt 2011). As indicated in Figure 5, the maximum curvature appears between the third and the seventh data points. In real-world applications the 'elbow point' cannot always be unambiguously identified (Ketchen and Shook 1996), and the choice of criteria depends on practical needs. Mathematically, the second derivatives of a curve show the slope of tangent line changes at each local point, which provides a useful indicator for 'the rate of change'. Here we define the 'Elbow point' as the point where the second derivative reaches the largest value between point 3 and point 7.

Figure 5 demonstrates that the Elbow value occurs when the number of clusters equals 5; therefore we classify the sample into five clusters. We then plot the point density distribution of user trajectories within the five clusters. As shown in Figure 6, these clusters are spatially distributed as follows:

Cluster 1: Clustered in the Northwest of the study area
Cluster 2: Evenly distributed across the city center
Cluster 3: Clustered in the North of the study area
Cluster 4: Clustered in the East of the study area
Cluster 5: Clustered in the Southwest of the study area (distant from the city center).

To better interpret the activity patterns of the five clusters, we also conducted a stop extraction based on the methdology described in Phithakkitnukoon *et al.* (2010).[2] We extracted the most frequent stops during daytime (7 am to 7 pm) on weekdays (these locations can be considered as work locations for most individuals working on a regular shift). Due to data resolution issues, we managed to extract the daytime stop locations for 584 users out of 1000. Figure 7 presents the centroids of these stops for each of the five clusters. This result confirms the differeces in spatial distribution of activity region for the five clusters in the previous Figure 6.

Investigating the spatial distribution of different population clusters can provide input for various types of social studies such as spatial segragation and neighborhood approximation. Besides the spatial distribution of activity space, previous research also focused

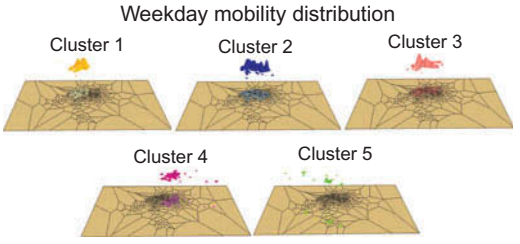

Figure 6. Point density distribution of the five clusters.

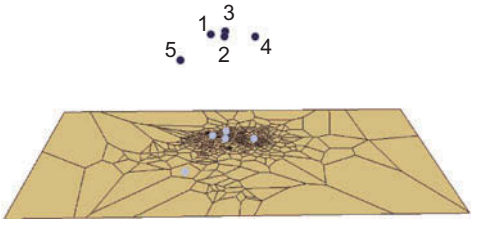

Figure 7. Centroids of most frequent stops during daytime for five clusters.

on investigating direction distribution of trips (Brockmann and Theis 2008). Since the movement between two consecutive phone calls can be considered as a displacement (Kang *et al*. 2012), we calculate the direction distribution of displacements in user trajectories and plot the direction distribution of these five clusters (the dotted red line indicates an average distribution of all five clusters).

The direction distribution provides valuable insights regarding the diversity of user movement. As indicated in Figure 8, all five clusters show a similar non-uniform pattern (in general higher distribution in the east–west direction compared to the north–south direction). More specifically, Clusters 1, 2, 3, and 5 appear to have two major directions (northeast east (NEE) and southwest west (SWW)), whereas Cluster 4 dominates on straight east (E) and west (W) directions. In traditional assumptions of Lévy flight, movement direction is considered as randomly distributed (Brockmann and Theis 2008); however, Liu *et al*. (2012) proved that the direction distribution of human movement is strictly restricted by the geographic boundary of the living environment. As shown in Figure 9, in City *A* the dominating direction of the urban area is NEE–SWW. This result confirms the findings in Liu *et al*. (2012) that the movement directions of human beings are restricted by the built-in environment, hence, they are not randomly distributed.

The Hellinger coefficient is often used to measure the correlation between two distributions (Vegelius *et al*. 1986). Given two concrete density distributions $p(x)$ and $q(x)$ defined on the same domain X, the Hellinger coefficient R_{H} ($R_{\mathrm{H}} \in [0,1]$) is given by the following equation:

$$R_{\mathrm{H}} = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{4}$$

Table 2 represents the Hellinger coefficient between the direction distribution of each cluster and the average distribution (the red dotted curves in Figure 8). As can be seen, the

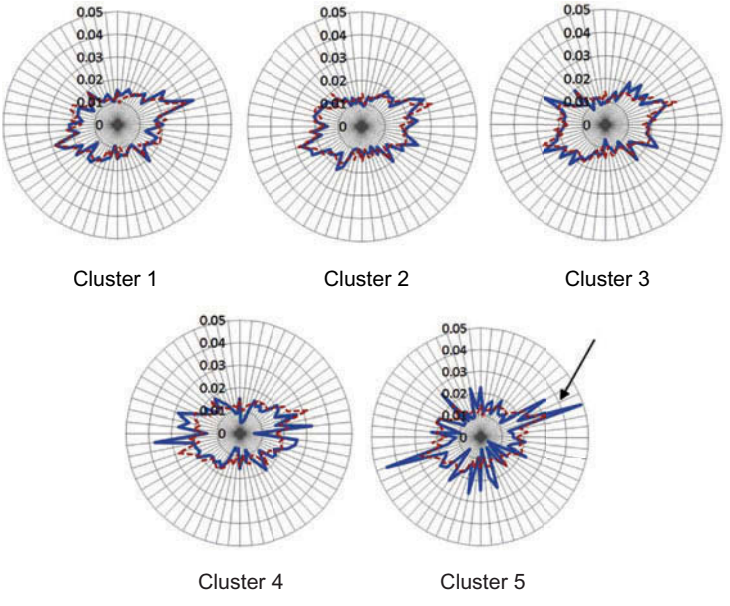

Figure 8. Direction distribution of the five clusters.

Figure 9. The shape of City *A* and its street networks.

coefficient between every pair is greater than 0.95, indicating a high level of similarity in terms of movement direction for the five clusters.

In addition, it is also useful to visually identify outlier patterns that deviate from regular circumstances. For instance, in Cluster 5 there appears to be a higher density of movements along the SWW–NEE direction compared to Clusters 1, 2, and 3.

In this section, we demonstrated the analysis results based on the proposed Spatio-temporal Edit Distance method. Another valuable part of the presented method is that it can be further customized based on context and practical needs. For instance, if researchers are interested in the question 'whose night-hour activity pattern is similar to User *A*?' they can incorporate the temporal constraints when calculating the distance between two trajectories. This category of time-related research questions will be discussed in the following section.

Table 2. Hellinger coefficients of direction distribution.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
R_H (compared to average pattern)	0.994	0.994	0.991	0.976	0.982

4.2. *Extended analysis including time dimension*

As discussed in Section 2, time plays an important role in measuring the similarity between two trajectories. For the time dimension, the semantic of the centroid can be considered as a central time point when a specific user prefers to make phone calls. Compared to the original Edit Distance method, it allows for calculating the displacement in the time dimension. Although the Spatio-temporal Edit Distance algorithm as demonstrated in Section 4.1 takes into account the order of how locations are visited, it does not explicitly represent the impact of time. In practice, it is highly possible that a comparison between trajectories involves constraints on when the activities are conducted. In this section we provide two analyses for comparing trajectories under a temporal constraint. The analyses are conducted from two perspectives: (1) Time as a cost function parameter; and (2) Time as a constraint in partitioning trajectories.

4.2.1. *Time as a cost function parameter*

Figure 10 shows the result of an analysis with $c = 0.5$ (cf., Equations (1), (2), and (3)). The example uses the weekend trajectory of the same User *A* as in Figure 2, where trajectory *D* shows the most similar user path when considering an equal contribution from both temporal and spatial dimensions. Figure 10b presents the distribution of phone calls at different times on 4 weekend days, where each point represents a recorded phone connection. When setting $c = 1$ (purely temporal) or $c = 0$ (purely spatial), the calculated Spatio-temporal Edit Distances between *A* and *D* are still smaller than 85% of the users in the sample, indicating that these two users demonstrate similar activity patterns in both space and time dimensions. We also added an explanatory sensitivity test of c values in the Appendix. In real-world applications, researchers can adjust the value of c based on practical needs.

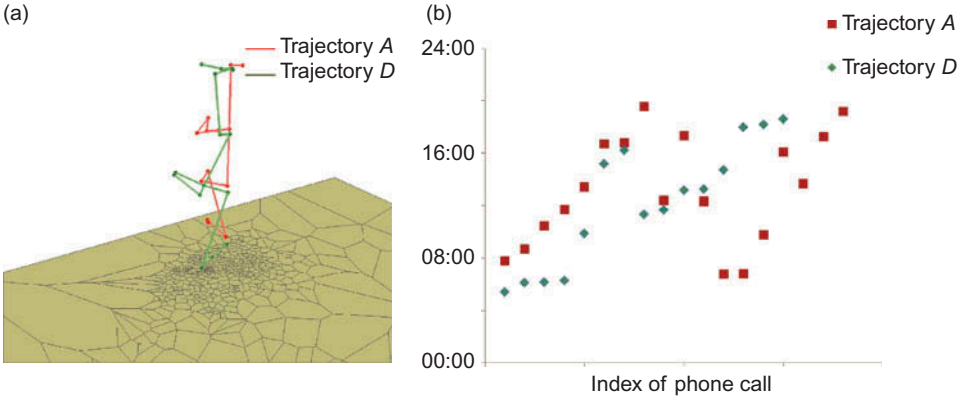

Figure 10. Two similar trajectories (a) space-time paths; (b) temporal patterns.

4.2.2. Time as a constraint in partitioning trajectories

The time-enabled cost function in Section 4.2.1 allows for explicitly investigating the influence of temporal factors when comparing two trajectories; in practice, users are often more interested in a certain time interval, such as 'does User *B* have a similar mobility pattern to User *A* in the morning/afternoon/evening?' In order to explore this type of questions, we extend the analysis from a different perspective, where time constraints are used to partition the trajectories; i.e., each trajectory is divided into sub-trajectories based on their timestamps in CDRs. To investigate the similarity of two trajectories under given temporal restrictions, it is necessary to calculate the distances between all pairs of sub-trajectories and construct a distance matrix which shows the correlation between each part of the trajectories. Here we demonstrate how this matrix is constructed and which activity patterns can be explored based on these matrices.

Given two users with trajectories R_1 $[p_{11}(x_{11}, y_{11}, t_{11}), \ldots, p_{1n}(x_{1n}, y_{1n}, t_{1n})]$ and R_2 $[(p_{21}(x_{21}, y_{21}, t_{21}), \ldots, p_{21}(x_{2m}, y_{2m}, t_{2m})]$ with n and m points, respectively, both R_1 and R_2 are regrouped into 4 sub-trajectories: $[R_{11}, R_{12}, R_{13}, R_{14}]$ and $[R_{21}, R_{22}, R_{23}, R_{24}]$, where each sub-trajectory R_{1j} (R_{2j}) only contains space-time points that fall into the corresponding time interval T_j ($j \in [1, 2, 3, 4]$), where T_j is defined as:

T_1 (Midnight): [0:00–6:00]; T_2 (Morning): [6:00–12:00];
T_3 (Afternoon): [12:00–18:00]; T_4 (Evening): [18:00–24:00]

The distance matrix is constructed as: D_{ij} = EditDistance(R_{1i}, R_{2j}), $i, j \in [1, 2, 3, 4]$. Note that R_{1i}, R_{2j} can be empty sets if the user does not have any call activity during the given period. In this case we manually assign that D_{ij} equals –1.

To illustrate the method, we further select 246 users from the previous 1000-user sample whose weekend trajectories have more than five time points in each of the last three time intervals (except for T_1 [0:00–6:00] due to low phone call counts after midnight). It is feasible to construct a matrix which indicates the (5%, 95%) range of calculated distances between different time intervals (Table 3). For instance, the pair (0.0282, 0.195) indicates that 5% of the distances between two trajectories in T_1 are smaller than 0.0282, while 95% of these distances are smaller than 0.195. This range table provides an effective way to conduct an exploratory analysis based on the sub-trajectories corresponding to different time intervals, as well as detecting potential outlier patterns.

Here we construct the distance matrix between two example Users *E* and *F* (Table 4).

Table 3. Range of distance matrix for different time intervals.

	T_1 Midnight	T_2 Morning	T_3 Afternoon	T_4 Evening
T_1 Midnight	(0.0282, 0.195)	(0.0543, 0.639)	(0.0548, 0.613)	(0.0437, 0.357)
T_2 Morning	–	(0.0382, 0.330)	(0.0379, 0.316)	(0.0361, 0.321)
T_3 Afternoon	–	–	(0.0373, 0.303)	(0.0357, 0.308)
T_4 Evening	–	–	–	(0.0322, 0.221)

Table 4. Example distance matrix.

	R_{e1}	R_{e2}	R_{e3}	R_{e4}
R_{f1}	0.128	**0.852**	0.358	0.341
R_{f2}	0.131	0.113	0.0567	**0.0342**
R_{f3}	0.0640	0.171	0.0702	0.0628
R_{f4}	0.0766	0.155	0.0589	0.0508

From Table 4, it is straightforward to identify the most similar sub-trajectory for User E in each time interval (marked as grey): $R_{e1} \rightarrow R_{f3}$; $R_{e2} \rightarrow R_{f2}$; $R_{e3} \rightarrow R_{f2}$; $R_{e4} \rightarrow R_{f2}$. The results demonstrate that the morning pattern (6:00–12:00) of User F is similar to the morning/afternoon/evening patterns of User E, whereas the afternoon pattern (12:00–18:00) of User F is similar to the midnight pattern of User E.

We can also extract outliers based on the reference ranges in Table 3. Here all the numbers that lie outside the ranges (Table 3) are considered outliers. As can be seen, two pairs of result distances (R_{e2}, R_{f1}) and (R_{e4}, R_{f2}) are recognized as outliers, indicating that the morning pattern of User E is different from the midnight pattern of User F (distance = 0.852; larger than 95% of the values in the sample), whereas the evening pattern of User E is very similar to the morning pattern of User F (distance = 0.0342, smaller than 5% of the values in our sample).

This section presented two extended analyses which incorporate the time dimension when calculating the similarity between trajectories. The method can be applied to various scenarios where researchers are interested in the impact of time when comparing different mobility patterns. Note that in this section we only demonstrated the method using individual examples; however, all the demonstrated analysis (i.e., clustering) in Section 4.1 can also be applied to explore the group patterns of user trajectories after considering the role of the time dimension.

5. Discussion

As presented in Sections 4.1 and 4.2, the Extended Edit Distance method can be used to measure the similarity of mobile user trajectories from CDRs. We demonstrated three types of analysis for this novel extension, where each analysis corresponds to a different scenario:

- *Edit Distance with spatial information*: This method calculates the cost function of each operation based on the displacement of the trajectory centroid ($c = 0$). Although the method considers the relative order of how the points are visited (i.e., site X is visited before site Y), it does not explicitly state the effect of absolute time (i.e., whether a site was visited in the morning or in the afternoon). This method can be used when time is not considered a major component in the analysis.
- *Edit Distance with spatial information and time as a parameter in the cost function*: This analysis includes time as a third dimension in the cost function. The parameter c controls the weight for both space and time dimensions. When $c = 1$ the cost function only considers the temporal deviation of points. This method is appropriate for analyses that test the overall effect of time.
- *Edit Distance with spatial information and time as a constraint in partitioning trajectories*: In this analysis, trajectories are divided into sub-components based on the time when each site is visited. From a technical perspective, this analysis uses the same distance function as in Section 4.1 but with a finer temporal resolution. It can be applied when researchers are interested in investigating patterns for different time intervals. The granularity of time intervals can be determined based on context and practical needs. This method works best when the phone records of a particular user are distributed evenly in time, which makes it feasible to extract sub-trajectories for each individual time interval.

From an empirical perspective, the analysis in Section 4.1 also indicates interesting patterns in the study area where trajectories are classified into five clusters. Each cluster concentrates on a specific region of the study area. There are no substantial differences in the moving direction analysis, i.e., all five clusters show a higher distribution in the east–west direction compared to the north–south direction, which is highly correlated with the layout of the study area; however, the results indicate a few abnormal patterns, for instance, Cluster 5 appears to have a higher number of movements in the SWW–NEE direction. As indicated in Figure 6, the point density distribution of Cluster 5 also deviates from the other four clusters, which provides valuable input for identifying users who behave differently from the majority (i.e., outlier patterns in Figure 4). Moreover, weekend patterns show a slightly smaller average distance (less diverse) than weekday patterns as indicated in Section 4.1, which is consistent with conclusions from previous research (Yuan *et al*. 2012). Our preliminary hypothesis is that in the study area most people still tend to visit regular locations (e.g., preferred grocery stores) in their leisure time, although movements on weekends have less spatial constraints (e.g., work locations), but more comprehensive data are needed to test this hypothesis, which is not the major focus of this paper.

In this analysis we adopted centroids as reference points when calculating the cost function. There are two major reasons for utilizing centroids in the proposed algorithm. First, as stated in Section 3.2, a centroid can be considered as the center of an activity region, which is often used as a reference or benchmark point in spatial point pattern analysis. Even though human beings may have more than one major activity region (e.g., home, work, gym, grocery stores), the centroid can still be used to represent the geometric center of the entire trajectory. This is not only applicable for monocentric movement patterns, but can also be applied to more complex scenarios. For instance, in motion detection and video surveillance studies, the centroid pixel is often selected as a representative pixel when tracking multicentered moving objects (Hu *et al*. 2004, Fernández-Caballero 2005). Second, the main point of developing the Spatial-temporal Edit Distance is to quantify the impact of each operation based on its spatial characteristics. For centroids, this impact can be easily determined by the displacement of the centroid after each deletion/insertion/replacement operation; therefore, it is important for the Extended Edit Distance method to have a single point of reference instead of different points in terms of calculating the cost function.

However, it is still interesting to investigate how other forms of reference points affect the cost function, such as utilizing regularly visited stops. These techniques may be helpful for improving the accuracy of the analysis in some cases. We provide a pilot study in the Appendix to compare the cost functions based on centroids and extracted stops. The results of this study indicate no significant differences. In addition, CDR data usually suffer from low resolution and accuracy; therefore, the procedure of extracting stops cannot be conducted for every user in the data set; instead, it only works when there are a sufficient number of records for a specific user. Hence, stop-based cost functions are not very flexible and generalizable for different data sets. The calculation of stops also increases the complexity of the methodology, which may be time consuming for large data sets (i.e., with terabytes of data). Based on above reasons, we believe that using centroids provides more flexibility for our data in this study, and allow us to make use of a larger portion of the information provided in the sparse data set.

In addition, the Edit Distance algorithm belongs to the family of dynamic programming, which is relatively time consuming with a time complexity of $O(m*n)$ (where m and n are the lengths of two series). For this research we used a study time interval of 9 days, which is relatively short. In order to compare data for a longer time interval, it will be

necessary to simplify the trajectories before conducting the distance calculation (i.e., aggregate points that are near to each other) to reduce computing complexity. It is also feasible to further improve the algorithm, such as define a binary or categorical cost function (i.e., two points within a threshold distance x are defined to be the same) if the application does not require an accurate numerical cost function. In addition, adopting high performance computing such as cluster computing or GPU computing can also be helpful for improving the performance of the algorithm.

It is also worth noting the unavoidable uncertainty issues when analyzing user trajectories utilizing CDR data. There are various types of uncertainty involved in our analysis. As argued in Xia (2005), uncertainty exists in the data sets to be mined, the mined knowledge and the process of applying new knowledge to other data. The major types of uncertainty involved in this analysis are (Yuan *et al*. 2012):

- *Low data quality due to insufficient knowledge*: The trajectories extracted from CDRs are both inaccurate and imprecise. The accuracy of data depends on the spatial density of base towers in the study area. In addition, the temporal resolution of data also depends on the frequency of phone calls; hence, as discussed in Section 1, CDRs can only represent an approximation of the real trajectories.
- *Imperfection of models and algorithms*: As stated in Box and Draper (1987, p. 424): 'Essentially, all models are wrong, but some are useful.' Although we have demonstrated the effectiveness of Spatio-temporal Edit Distance method when comparing human trajectories, using alternative methods will inevitably impact the uncertainty of the results (i.e., the choice of different cost functions).
- *Natural variability of human mobility*: Although previous research has proved the predictability of human mobility (González *et al*. 2008), randomness is an essential nature of human mobility.

Several potential methods can be adopted to quantitatively measure these uncertainty issues, such as probability theories, Bayesian network, and fuzzy sets. This also provides an important direction for future research.

6. Conclusions and future work

In this paper, we have developed a Spatio-temporal Edit Distance method for measuring trajectory similarity of mobile phone users. The main contribution of this paper is the cost function to extend the traditional Edit Distance algorithm in order to incorporate both spatial and temporal factors. We also demonstrated its effectiveness in case studies based on a sample data set. Three types of analysis were presented for scenarios with and without the effect of time: (1) Edit Distance with spatial information; (2) Edit Distance with time as a factor in the cost function; and (3) Edit Distance with time as a constraint in partitioning trajectories. The outcomes of this research contribute to both methodological and empirical perspectives. Compared to closely related work by Shoval and Isaacson (2007), our study advances this research from the following perspectives:

- Incorporated spatial coordinates in cost functions. From a statistical perspective, there are four levels of measurement: *Nominal*, *Ordinal*, *Interval*, and *Ratio* (Stevens 1946). Shoval and Isaacson (2007) incorporated nominal measurement (categorical geographic locations such as 'home' and 'work') in sequence

alignment to study human activity patterns; whereas in this research, we take a step further by adding ratio measurement (true space-time coordinates) in measuring the similarity of two sequences.
- Tested the method in a large-scale environment (approximately 450 km^2) compared to their small-scale environment (Akko's old city, approximately 0.5 km^2).

The method can be applied for exploring both individual-level and group-level patterns of user trajectories. Potential cases for analysis include but are not limited to:

- Individual-level: This method may be used to identify similar trajectory patterns for a given user. Computation can be from a spatial, temporal, and spatio-temporal perspective. It is also feasible to explore the similarity/dissimilarity between different sub-trajectories of two users based on predefined time intervals.
- Group-level: For a group of users, the Spatio-temporal Edit Distances can be employed to construct a distance matrix, which can be later used for user trajectory clustering and outlier pattern detection. It is also interesting to calculate the average distance between each user and the others, and use information to extract outlier users who behave differently from the majority.

The analysis of the sample data set revealed several remarkable trends and characteristics of the study area. The average distances of users follow a skewed normal distribution, which indicates that there are several outlier patterns of users that behave differently compared to the other users in the sample set. It was discovered that these outliers have large activity spaces in and out of the city. The five clusters demonstrate the major movement direction in the city (east–west) with certain outlier patterns (i.e., Cluster 5).

This research provides us with new insights regarding the study of similarity between two trajectories based on CDRs. The analysis demonstrated that the proposed algorithm can be easily applied to identify patterns in a cellular environment, as well as providing input for policy-makers and the location-based services market. The proposed method also contributes to the advancement of geographic knowledge discovery in the age of instant access, where trajectory data are commonly acquired with low resolution and precision.

In the future we will further modify the algorithm to incorporate various types of cost functions that also take into account the movement directions and additional factors/dimensions of importance, such as location semantics (i.e., home, work). This is necessary since two trajectories with the same geometric properties can be considered different if they are different in other aspects (i.e., with different travel purposes). A thorough understanding of these factors is essential for designing appropriate cost functions for practical needs. In addition, when eliminating redundant points, we predefined the threshold as 0.5 hour, it may be worthwhile to investigate how this threshold influences the performance of the algorithm. One limitation of this method is the lack of reliability assessment (Shoval and Isaacson 2007); therefore, in the future it will be necessary to develop a framework for quantifying its effectiveness and the uncertainty issues. We will also look into applying the method to other cities and countries to test its robustness. Moreover, in this research we take the average value of EditDistance(A, B) and EditDistance(B, A) to keep the symmetry of the measurement; it will be interesting to investigate how asymmetry affects the results. Another promising future work direction relates to comparing the results of Edit Distance for CDRs and regular trajectory similarity measures for GPS data. This method can also be utilized in a wide range of application fields, including crime pattern detection and traffic analysis.

Acknowledgments

This research is supported by the Swiss National Science Foundation (Project # 205121_141284). We thank Dr Yu Liu and Geosoft Lab at Peking University for providing the data and insightful comments at an earlier stage of this work. The three reviewers provided excellent feedback, which helped us to improve the content and clarity of this paper.

Notes

1. The city name is not shown as required by the data provider.
2. In Phithakkitnukoon *et al.* (2010), the trajectories are regrouped into sub-trajectories based on the restriction that any two consecutive points within a sub-trajectory are located within the cell of the same mobile phone tower. If the time duration of a sub-trajectory is longer than the temporal threshold ΔT (here defined as 0.5 hour), the sub-trajectory is identified as a stop for the particular user. Here we conduct the extraction process on the original trajectories (before removing redundant points as discussed in Section 3.2).

References

Abbott, A., 1995. Sequence-analysis – new methods for old ideas. *Annual Review of Sociology*, 21, 93–113.

Ahas, R., 2005. Mobile phones and geography: social positioning method. *In*: *Power over time-space: inaugural Nordic geographers meeting*, 10–14 May, Lund, 1–8.

Ahas, R., *et al.*, 2010. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. *Transportation Research Part C – Emerging Technologies*, 18 (1), 45–54.

Baert, A.E. and Seme, D., 2004. Voronoi mobile cellular networks: topological properties. *In*: *ISPDC 2004: Third International Symposium on Parallel and Distributed Computing/ Heteropar '04: Third International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks*, *Proceedings*, 5–7 July, Cork, 29–35.

Bagrow, J.P. and Koren, T., 2009. Investigating bimodal clustering in human mobility. *In: International Conference on Computational Science and Engineering*, 29–31 August, Vancouver. Washington, DC: IEEE Computer Society, 944–947.

Baron, R., 2012. *Computational drug discovery and design*. New York: Humana Press.

Box, G.E.P. and Draper, N.R., 1987. *Empirical model-building and response surfaces*. New York: Wiley.

Brockmann, D. and Theis, F., 2008. Money circulation, trackable items, and the emergence of universal human mobility patterns. *IEEE Pervasive Computing*, 7 (4), 28–35.

Buchin, K., *et al.*, 2009. Finding long and similar parts of trajectories. *In*: O. Wolfson, D. Agrawal, and C.-T. Lu, eds. *17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*, 4–6 November, Seattle WA. Chicago, IL: ACM, 296–305.

Chen, L., Özsu, M.T., and Oria, V., 2003. *Robust and efficient similarity search for moving object trajectories*. Technical Report CS-2003-30. School of Computer Science, University of Waterloo.

Dodge, S., Laube, P., and Weibel, R., 2012. Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science*, 26 (9), 1563–1588.

Du Mouza, C. and Rigaux, P., 2005. Mobility patterns. *Geoinformatica*, 9 (4), 297–319.

Eagle, N., Pentland, A., and Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (36), 15274–15278.

Eiter, T. and Mannila, H., 1994. *Computing discrete Fréchet distance*. Technical Report CD-TR 94/ 64. Vienna: Christian Doppler Laboratory for Expert Systems, Technical University of Vienna.

Everitt, B., 2011. *Cluster analysis.* 5th ed. Chichester, West Sussex: Wiley.

Fernández-Caballero, A., 2005. Motion direction detection from segmentation by LIAC, and tracking by centroid trajectory calculation. *In*: *The Fifth International Workshop on Pattern Recognition in Information Systems*, 24–25 May, Miami, FL.

Foss, A. and Zaïane, A., 2002. A parameterless method for efficiently discovering clusters of arbitrary shape in large datasets. *IEEE International Conference on Data Mining (ICDM)*, 9–12 December, Maebashi. Washington, DC: IEEE Computer Society, 179–186.

González, M.C., Hidalgo, C.A., and Barabasi, A.L., 2008. Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–782.

Han, M., *et al*., 2004. An algorithm for multiple object trajectory tracking. *In*: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol 1, 27 June–2 July, Washington, DC, 864–871.

Hu, W., *et al*., 2004. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34 (3), 334–352.

Janowicz, K., Raubal, M., and Kuhn, W., 2011. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2, 29–57.

Joh, C.H., *et al.*, 2002. Activity pattern similarity: a multidimensional sequence alignment method. *Transportation Research Part B – Methodological*, 36 (5), 385–403.

Johnson, R.A., 2007. *Advanced Euclidean geometry.* Mineola, TX: Dover Publications.

Kang, C., *et al.*, 2012. Intra-urban human mobility patterns: an urban morphology perspective. *Physica A: Statistical Mechanics and Its Applications*, 391 (4), 1702–1717.

Kang, H.-Y., Kim, J.-S., and Li, K.-J., 2009. Similarity measures for trajectory of moving objects in cellular space. *In*: S.Y. Shin and S. Ossowski, eds. *SAC*, 9–12 March, Honolulu, HI. ACM, 1325–1330.

Ketchen, D.J. and Shook, C.L., 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17 (6), 441–458.

Kolbe, T.H., Becker, T., and Nagel, C., 2008. *1st Technical Report – Discussion of Euclidean Space and Cellular Space and Proposal of an Integrated Indoor Spatial Data Model*. Berlin: Technische Universität Berlin.

Kwan, M.P., 2000. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C – Emerging Technologies*, 8 (1–6), 185–203.

Kwan, M.P., 2004. GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns. *Geografska Annaler B*, 86 (4), 267–280.

Kwan, M.P., 2006. Transport geography in the age of mobile communications. *Journal of Transport Geography*, 14 (5), 384–385.

Larsen, J., Urry, J., and Axhausen, K.W., 2006. *Mobilities, networks, geographies*. Aldershot: Ashgate.

Lee, J.-G., Han, J., and Whang, K.-Y., 2007a. Trajectory clustering: a partition-and-group framework. *In*: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 12–14 June, Beijing. ACM, 593–604.

Lee, H.J., Lee, Y.J., and Lee, C.W., 2001. Gesture classification and recognition using principal component analysis and HMM. *In*: H.-Y. Shum, M. Liao, and S.-F. Chang, eds. *Advances in Multimedia Information Processing – PCM 2001 Proceedings*, 24–26 October, Bejing, Volume 2195 of Lecture Notes in Computer Science. Berlin: Springer, 756–763.

Lee, J., Rajauria, P., and Shah, S.K. 2007b. A model-based conceptual clustering of moving objects in video surveillance. *In*: *Proceedings of SPIE IS&T electronic imaging*, 28 January–1 February 2007, San Jose, CA.

Liu, Y., *et al.*, 2012. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14 (4), 463–483.

Long, J.A. and Nelson, T.A., 2013. A review of quantitative methods for movement data. *International Journal of Geographical Information Science*, 27 (2), 292–318.

Maier, D., 1978. The complexity of some problems on subsequences and supersequences. *Journal of ACM*, 25 (2), 322–336.

Makinen, V., 2001. Using edit distance in point-pattern matching. *In*: *Eighth Symposium on String Processing and Information Retrieval*, *Proceedings*, 13–15 November, Laguna de San Rafael. IEEE, 153–161.

Marwan, M., Fuad, M., and Marteau, P.F., 2008. The extended edit distance metric. *In*: *2008 International Workshop on Content-Based Multimedia Indexing*, 18–20 June, London. IEEE, 226–232.

Miller, H.J., 2005. A measurement theory for time geography. *Geographical Analysis*, 37 (1), 17–45.

Miller, H.J., 2009. Geographic data mining and knowledge discovery: an overview. *In*: H.J. Miller and J. Han, eds. *Geographic data mining and knowledge discovery* 2nd ed. London: CRC Press, 3–32.

Miller, H.J. and Han, J., 2009. *Geographic data mining and knowledge discovery*. 2nd ed. Boca Raton, FL: CRC Press.

Phithakkitnukoon, S., *et al.*, 2010. Activity-aware map: identifying human daily activity pattern using mobile phone data. *In*: A.A. Salah, *et al*., eds. *HBU 2010.* Heidelberg: LNCS, Springer, 14–25.

Potamias, M., Patroumpas, K., and Sellis, T., 2006. Sampling trajectory streams with spatiotemporal criteria. *In*: *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*, 3–5 July, Vienna. IEEE, 275–284.

Powell, D.R., Allison, L., and Dix, T.I., 2000. Fast, optimal alignment of three sequences using linear gap costs. *Journal of Theoretical Biology*, 207 (3), 325–336.

Rucklidge, W.J., 1997. Efficiently locating objects using the Hausdorff distance. *International Journal of Computer Vision*, 24 (3), 251–270.

Salvador, S. and Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *In*: *ICTAI 2004: 16th IEEE International conference on Tools with Artificial Intelligence*, *Proceedings*, 15–17 November, Boca Raton, FL. Los Alamitos, CA: IEEE Computer Society Press, 576–584.

Schneider, M., 1999. Uncertainty management for spatial data in databases: fuzzy spatial data types. *Advances in Spatial Databases*, 1651, 330–351.

Senin, P., 2008. *Dynamic time warping algorithm review*. Honolulu, HI: University of Hawaii at Manoa.

Sharifzadeh, M. and Shahabi, C., 2006. Utilizing Voronoi cells of location data streams for accurate computation of aggregate functions in sensor networks. *Geoinformatica*, 10 (1), 9–36.

Shoval, N. and Isaacson, M., 2007. Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers*, 97 (2), 282–297.

Smith, T.F. and Waterman, M.S., 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147 (1), 195–197.

Song, C.M., *et al.*, 2010. Limits of predictability in human mobility. *Science*, 327 (5968), 1018–1021.

Stergiopoulos, J. and Tzes, A., 2009. Voronoi-based coverage optimization for mobile networks with limited sensing range – a directional search approach. *In*: *2009 American Control Conference*, Vols 1–9, 10–12 June, St Louis, MI. IEEE, 2642–2647.

Stevens, S.S., 1946. On the theory of scales of measurement. *Science*, 103 (2684), 677–680.

Tibshirani, R., Walther, G., and Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B – Statistical Methodology*, 63, 411–423.

Tversky, A., 1977. Features of similarity. *Psychological Review*, 84 (4), 327–352.

Vegelius, J., Janson, S., and Johansson, F., 1986. Measures of similarity between distributions. *Quality & Quantity*, 20 (4), 437–441.

Wagner, R.A. and Fischer, M.J., 1974. The string-to-string correction problem. *Journal of the ACM*, 21 (1), 168–173.

Ward, J.H., 1963. Hierarchical rouping to optimize an objective function. *Journal of the American Statistical Association*, 58 (301), 236–244.

Wilkes, M., 2008. *A graph-based alignment approach to context-sensitive similarity between climbing routes*. (Master Thesis). University of Münster.

Won, J.I., *et al*., 2009. Trajectory clustering in road network environment. *In*: *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 30 March–2 April, Nashville, TN. IEEE, 299–305.

Xia, Y., 2005. *Integrating uncertainty in data mining*. Doctoral dissertation. University of California.

Xia, Y., *et al.*, 2011. Spatio-temporal similarity measure for network constrained trajectory data. *International Journal of Computational Intelligence Systems*, 4 (5), 1070–1079.

Yang, M., Kpalma, K., and Ronsin, J., 2008. A survey of shape feature extraction techniques. *In*: P.-Y. Yin, ed. *Pattern Recognition.* Rijeka: InTech, 43–90.

Yang, M., Kpalma, K., and Ronsin, J., 2012. Shape-based invariant feature extraction for object recognition. *In*: R. Kountchev and K. Nakamatsu, eds. *Advances in reasoning-based image*

processing intelligent systems: conventional and intelligent paradigms. New York: Springer, 255–314.
Yang, S. and Tiow, T.T., 2007. Long distance redundancy reduction in thin client computing. *In*: *6th IEEE/ACIS International Conference on Computer and Information Science*, *Proceedings*, 11–13 July, Melbourne. Los Alamitos, CA: IEEE Computer Society, 961–966.
Yuan, Y. and Raubal, M., 2012a. Extracting dynamic urban mobility patterns from mobile phone data. *In*: N. Xiao, *et al*., eds. *Geographic Information Science – 7th International Conference (GIScience 2012)*, 18–21 September, Columbus, OH, 354–367.
Yuan, Y. and Raubal, M., 2012b. Similarity measurement of mobile phone user trajectories – a modified edit distance method. *In*: *Workshop on "Progress in Movement Analysis – Experiences with Real Data"*, 15–16 November, Zurich, Switzerland.
Yuan, Y., Raubal, M., and Liu, Y., 2012. Correlating mobile phone usage and travel behavior – a case study of Harbin, China. *Computers, Environment and Urban Systems*, 36 (2), 118–130.
Zheng, Y. and Zhou, X., 2011. *Computing with spatial trajectories*. New York: Springer.
Zini, M., *et al*., 2006. Plagiarism detection through multilevel text comparison. *AXMEDIS 2006: Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*, *Proceedings*, 13–15 December, Leeds. Los Alamitos, CA: IEEE Computer Society, 181–185.

Appendix

A. Sensitivity test of *c* values.

Here we provide a sensitivity test for the constant *c*. For the same example User *A* in Figure 2, we calculate the average distance between *A* and the other 999 users in the sample set under different *c* values (see attached figure, $0 \leq c \leq 1$). As can be seen from Figure A1, when the value *c* increases, the average distance increases smoothly. This is helpful for understanding how the distance values vary for a single user when assigning different weights for the spatial and temporal dimension; however, for inter-trajectory analysis, the relative ranking plays a more important role than the absolute value, so the magnitude of the values under different *c* values cannot provide direct evidence when comparing between different users. In practice the choice of *c* should be determined by the specific requirements.

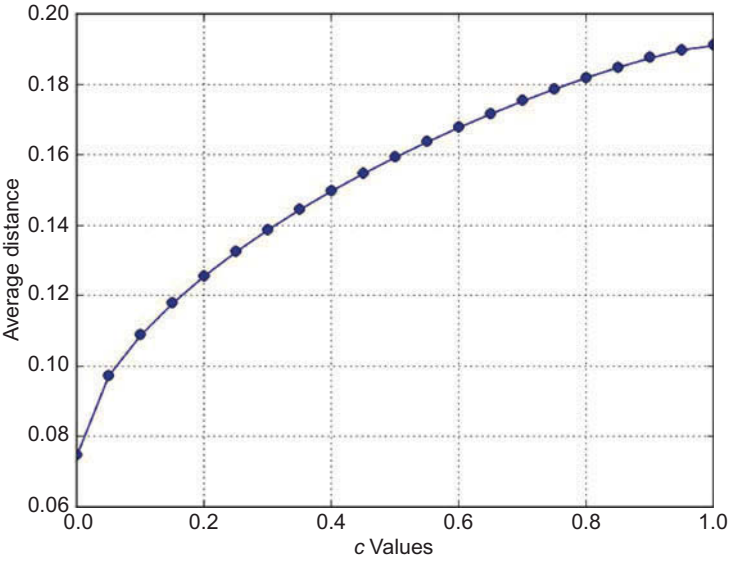

Figure A1. Sensitivity test of *c* values.

B. The choice of reference points in cost functions

We conducted a pilot study which utilizes estimated stops to calculate the cost function (as a comparison to the centroid-based algorithm). The cost function is described as follows:

- Estimate two major stops for each user based on the algorithm described in Section 4.1.2. As argued by Bagrow and Koren (2009), most human beings have at least two major regularly visited points, in most cases, home and work locations. Here we select the most w (Monday – Friday 8 am to 5 pm) and the most frequent stop during night hours (7 pm to 7 am).
- Redefine cost function

Cost [Delete (s_i)] = the average distance between s_i and the estimated two stops of trajectory S.
Cost [Insert (t_j)] = the average distance between t_j and the estimated two stops of trajectory S.
Cost [Replace (s_i, t_j)] = the distance between s_i and t_j.

- Case study

We then conducted an average Edit Distance analysis as in Figure 4. We selected another 1000 users for whom the most frequent stops are extracted for both day and night hours. Note that these are not the same 1000 users as in Section 4, due to the fact that the stop extraction only works for a small fraction of the users based on their CDR records; therefore we expanded the selected sample to ensure a 1000 sample size for the pilot study. Although the magnitude of the average distance changes, the result shows the ranking of these distances still appears to be stable, and no significant difference was found based on a Wilcoxon Signed Ranks Test for the difference in rankings. Table A1 also shows the top 20 users with the smallest average distances as a demonstration. Except for one user (highlighted in bold font) the ranking of average distances is the same as when using the original centroid-based method.

This pilot study indicates that it is a feasible option to use points of interest (POIs) as reference points in the cost functions. However, as discussed in Section 5, we believe that using the centroid is more flexible and reliable for our data in this study.

Table A1. Users with the smallest distances.

User ID	Average distance (Centroid-based method)	Ranking (Centroid-based method)	Average distance (Stop-based method)	Ranking (Stop-based method)
03643	0.09825689	1	0.558182	1
04519	0.099072215	2	0.628764	2
03636	0.099576107	3	0.752597	3
03659	0.099875414	4	0.779598	4
03637	0.100641924	5	0.823171	5
03670	0.100710201	6	0.853321	6
03689	0.100829396	7	0.879501	7
03601	0.101147964	8	0.927403	8
04502	0.101760268	9	0.942962	9
04841	0.102657226	10	0.9449	10
04511	0.102760537	11	0.94915	11
03660	0.103254855	12	0.962066	12
03663	0.103710112	13	0.965628	13
03606	0.103852192	14	1.05283	14
03607	0.105425101	15	1.163057	15
03636	0.105427964	16	1.199316	16
03664	0.106649642	17	1.370627	17
03623	0.106771588	18	1.404297	18
*****04815*****	**0.107221141**	**19**	**1.736047**	**23**
03658	0.107759978	20	1.634979	20