International Journal of Geographical Information Science
Vol. 32, No. 01, January 2018, 140–168

Research Article

Unveiling Movement Uncertainty for Robust Trajectory Similarity Analysis

Andre Salvaro Furtado[a,b*], Luis Otavio Campos Alvares[b], Nikos Pelekis[c], Yannis Theodoridis[c] and Vania Bogorny[b]

[a]*Instituto Federal de Santa Catarina, Xanxere, Brazil*; [b]*PPGCC, INE, Universidade Federal de Santa Catarina, Florianopolis, Brazil*; [c]*University of Piraeus, Piraeus, Greece*

(*v3.0 August 2017*)

Trajectory data analysis and mining require distance and similarity measures, and the quality of their results is directly related to those measures. Several similarity measures originally proposed for time-series were adapted to work with trajectory data, but these approaches were developed for well-behaved data, that usually do not have the uncertainty and heterogeneity introduced by the sampling process to obtain trajectories. More recently, similarity measures were proposed specifically for trajectory data, but they rely on simplistic movement uncertainty representations, such as linear interpolation. In this article we propose a new distance function, and a new similarity measure that uses an elliptical representation of trajectories, being more robust to the movement uncertainty caused by the sampling rate and the heterogeneity of this kind of data. Experiments using real data show that our proposal is more accurate and robust than related work.

Keywords: Movement Similarity, Raw Trajectory Similarity, Elliptical Trajectory Representation, Dynamic Threshold Similarity, Parameter free Similarity Measure

1. Introduction

The increasing use of GPS-enabled devices allowed the collection of huge amounts of data representing the movement history of individuals. When an individual is moving, its location is collected along time in the form of sequences of space-time points, called *raw trajectories*. A raw trajectory is a sequence $\langle p_1, \ldots, p_n \rangle$ of points $p = ((x, y), t)$, where (x, y) is the geographic position of the individual at the time instant t.

*Corresponding author. Email: asalvaro@inf.ufsc.br

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/13658816.2017.1372763
http://www.informaworld.com

Real trajectory data are in general collected over different sampling strategies, and this process introduces uncertainty in the movement representation (Pfoser and Jensen 1999). This uncertainty is caused by two types of error: i) the measurement error, caused by the impossibility to determine the position of an object by the measurement system that is intrinsic to each sampled point; and ii) the interpolation error, that refers to the limitations to represent the motion between two sampled points, influenced by the sampling rate (Ranacher *et al.* 2016). However the heterogeneous distribution of trajectory points still remains a problem in real trajectory datasets. A consequence is that two individuals moving in the same path may have different trajectories, making the movement similarity analysis a complex and challenging task.

Movement similarity analysis is useful for several application domains, such as: i) in a set of student trajectories it is possible to determine which students follow similar paths between their homes and the university, what is useful in a ride-sharing *app*; ii) in taxi trajectories, we may determine if a trajectory between two popular regions (e.g., airport and the city center) is very dissimilar to the others, to identify a driver that took a longer path to increase the fare; and iii) in animal trajectories we may identify the species through the most similar trajectories in a database where the species are already known.

Similarity measures have been proposed for different purposes in information retrieval and data mining, such as: top-K similarity queries - queries that given a trajectory return the most similar trajectories; trajectory outlier detection - identify the objects that move differently from the majority; and clustering techniques for grouping most similar trajectories, and classification of individuals by socio-demographic profiles (e.g., student, worker or retired) according to their trajectories

For similarity analysis there are the well-known DTW (Dynamic Time Warping) (Berndt and Clifford 1994), developed for time series, LCSS (Longest Common Subsequence) (Vlachos *et al.* 2002), EDR (Edit Distance on Real Sequences) (Chen *et al.* 2005), SWALE (Morse and Patel 2007), wDF (Ding *et al.* 2008), MSM (Furtado *et al.* 2016), and others. In general, these approaches have considered the physical properties of raw trajectories, and a summary of these measures is presented in Ranacher and Tzavella (2014). The main limitations of these approaches are related to the assumption that the movement between two points is a straight line (linear interpolation) or building a circle of fixed size around every *single* point for point matching (called matching threshold). When dealing with real trajectory datasets, where points are collected at different sampling rates, these limitations directly affect the accuracy in the similarity assessment, as detailed later in Section 2.

In this work we focus on the spatial similarity of movement, where two objects are considered similar if they share a similar path in space. As the granularity of the sampling rate may be variable and may have, for instance, minutes between two samples, building a radius of fixed size around each trajectory point for measuring spatial similarity is a very limited solution. To solve the problem, in this paper we introduce the idea of dynamic and automatic threshold definition, and use the concept of ellipses proposed by Pfoser and Jensen (1999), between two consecutive sampled points, in order to improve the accuracy in trajectory similarity analysis. We use an elliptical representation of trajectory, where the size of the ellipses is dynamically computed according to the distance between two consecutive trajectory points. This approach avoids the need of a fixed point threshold or linear interpolation, overcoming several problems. Considering this new representation, we propose a novel trajectory similarity measure and a new distance function to estimate a narrowed upper bound area for the ellipses. We show with extensive experiments that

our approach is more robust and precise than related similarity measures.

In summary, we make the following contributions:

(1) We propose a new distance function to estimate an approximate upper bound for the real movement between two points in space, and use this distance to dynamically determine the size of the matching threshold;
(2) We introduce a new parameter-free trajectory similarity measure, called UMS (Uncertain Movement Similarity) that is based on dynamically defined thresholds and of variable size, thus being more robust to variations in the sampling rate and more accurate in the similarity assessment.
(3) We perform extensive experiments based on state-of-the-art techniques used in the literature to show that our approach is more robust and accurate than related work. It includes the extension of the classic retrieval-based evaluation technique (precision @ recall) to work with trajectory data.

The remainder of this article is organized as follows: Section 2 presents the Problem Statement. Section 3 presents the basic concepts and the new distance function for the proposed similarity measure. Section 4 defines the UMS, a new similarity measure for trajectory data. Section 5 validates the proposed measure using a variety of experiments. Section 6 presents the related work and Section 7 remarks the main contributions and concludes the article.

2. Problem Statement

In this section we present the basic concepts and point out the main problems of existing approaches for trajectory similarity analysis. The real movement of an individual is continuous along time, as stated in Definition 2.1.

Definition 2.1: Movement. The movement of an individual is represented by a continuous function $\mathcal{M} : \mathrm{R}_+ \rightarrow \mathrm{R}^2$, assigning time instants over a two-dimensional space.

Although the real movement of an object is continuous, mobile devices collect discretized points along time, such that the movement is discretized as a sequence of timestamped spatial locations, as stated in Definition 2.2.

Definition 2.2: Raw Trajectory. A raw trajectory T is a sequence of time-ordered sampled points $\langle p_1, ..., p_n \rangle$, where each point has the form $p_k = ((x, y)_k, t_k)$, such that $(x, y)_k$ is the location in space and t_k is the time instant that $(x, y)_k$ was sampled.

Considering the definitions of movement and raw trajectory, the central problem of this work can be stated by the following question: Given two trajectories $R = \langle r_1, \ldots, r_n \rangle$ and $S = \langle s_1, \ldots, s_m \rangle$, how similar is their movement? In the following we describe three main problems that affect existing similarity measures:

Problem 2.3 Rigid Uncertainty Representation (RUR). *The sampling process introduces uncertainty in the representation of raw trajectories. To attenuate the effect of uncertainty, two techniques are commonly used over the sampled points: i) linear interpolation; and ii) point threshold.*

In approaches that use *linear interpolation*, such as wDF (Ding *et al.* 2008) or as adopted in EDwP (Ranu *et al.* 2015), the assumption is that the movement between two sampled points is a straight line, what is not the case in real movement, where individuals usually follow non-linear paths to avoid obstacles. Figure 1(a) illustrates an example of

three trajectories P, Q and R, where small circles correspond to the sampled points. Figure 1(b) illustrates the representation of P, Q and R using linear interpolation, with the movement between the sampled points assumed as a straight line. The methods that use linear interpolation will result in $similarity(P, R) > similarity(P, Q)$, while considering the real movement (in Figure 1(a)), clearly the $similarity(P, R) < similarity(P, Q)$.

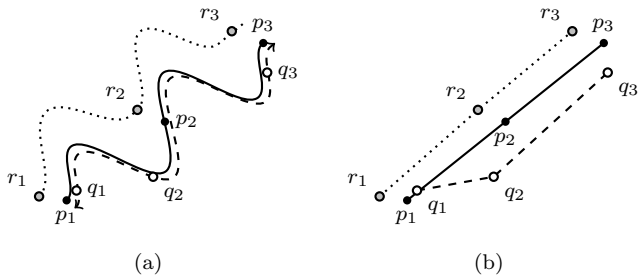


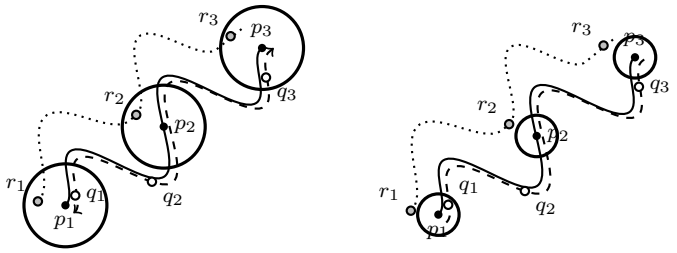
Figure 1. Real Movement (a) and Linear Interpolation (b) representation for trajectories P, Q and R (contiguous, dashed and dotted lines)

Approaches that use *point threshold*, such as LCSS (Vlachos *et al.* 2002), EDR (Chen *et al.* 2005), SWALE (Morse and Patel 2007), CATS (Hung *et al.* 2015), and MSM (Furtado *et al.* 2016), build a circle of radius with fixed size, given by the user, around each trajectory sampled point, assuming that the sampled points could be at any location inside the radius. An example is shown in Figure 2(a), where a circle of given fixed size is built around each point of trajectory P. There are two problems in this approach: first, the analysis occurs around every single sampled point of a trajectory; and second, a single threshold value cannot be accurately defined for trajectories with heterogeneous distribution of sampled points[1]. The methods that use point threshold to determine if two points of different trajectories are similar, verify if a point of a trajectory Q is within the radius around a point of a trajectory P, and if this is the case, they are said to match. Figure 2 (a) illustrates this problem where there are parts of the real movement of P that are not covered by the threshold radius around the points of P. As as consequence, although the real movement of trajectory Q crosses the radius around p_2, as there is no sampled point of Q inside the radius of p_2, there will be no match between p_2 and Q. Because of the threshold uncovered areas, the methods LCSS, EDR and CATS will not match p_2 with any point of Q, decreasing the similarity score, even though the real movement of P (between p_1 and p_2) was very similar to the real movement of Q. As a consequence, for these methods an unexpected result $similarity(P, R) > similarity(P, Q)$ is obtained. The expected result $similarity(P, Q) > similarity(P, R)$ can be obtained by reducing the threshold size, as shown in Figure 2(b). However, with a lower value for the point threshold the original problem not only remains but has its effect increased, losing the matching between points p_3 and q_3.

In addition, the definition of a fixed threshold value for all trajectories is not accurate in scenarios where the density distribution of points varies significantly. For instance, the density of a trajectory in low speed areas of a city will be higher than in a highway, what

[1]Heterogeneous distribution of sampled points is common in real-world data. It happens because the time-sampling strategy is commonly adopted (e.g., a point is collected each 30s). As a consequence of this strategy, when the moving speed of an individual is higher, the distance between its sampled points tend to be greater, resulting in a less dense distribution of points. For instance: consider an individual driving at 120km/h in a highway and then reaching the city center where the speed is reduced to 40km/h. Because of the higher speed the distance between sequential points in the first part is larger.

makes the distance between sampled points of the same trajectory to vary dramatically. Therefore, if the threshold is defined with a high value, it can match trajectories that performed different movements in nearby streets, while a low value can lead to none matching points of trajectories with similar movement and in the same street.



(a) Big Threshold (b) Small Threshold

Figure 2. Point Radius with large (a) and small (b) threshold values what directly influence the similarity score

Methods that directly calculate the distance between the sampled points such as DTW and wDF are also sensitive to sampling gaps. For example, in Figure 2(a), the sum of the point distances $dist(p_1, r_1) + dist(p_2, r_2) + dist(p_3, r_3)$ of trajectories P and R is smaller than the sum of the distances between P and Q $dist(p_1, q_1) + dist(p_2, q_2) + dist(p_3, q_3)$. As a result, the $similarity(P, R) > similarity(P, Q)$.

Problem 2.4 Binary Distance Assumption (BDA). *Some methods as LCSS and EDR use binary values to measure the similarity score between two points: if two points match, the score is 1, otherwise it is 0 (the inverse in EDR).*

The main problem of the binary distance assumption is that the real distance between the points inside the threshold is not considered, i.e., if the threshold of p is $100m$, a point of another trajectory that is $1m$ or $99m$ far from p will have the same score in relation to p. Figure 2(a) illustrates this problem for trajectories P, Q and R. Looking at the points p_1, q_1 and r_1, it is clear that $dist(p_1, q_1) < dist(p_1, r_1)$, but LCSS and EDR will consider the same distance between the points.

Problem 2.5 Sampling Rate Intolerance (SRI). *Let P and Q be two trajectories with very similar real movement but different sampling rates, as shown in Figure 3. Trajectory P has 3 points and trajectory Q has 5 points. Most related approaches are not tolerant to this situation. For example, for EDR, the maximum similarity score would be $sim(P, Q) \leq 0.6$ because the method matches points with a one-to-one cardinality, hence, at least two points of Q will not have any matching, therefore decreasing the similarity score.*

EDR looks for the best matching sequence in two trajectories, but in this process one point of a trajectory can match only one point of the other trajectory. As trajectory Q has more sampled points than trajectory P, an error is introduced $(length(Q) - length(P))$, decreasing the similarity. In this same case, DTW would repeat the nearest point of P to the non-matched point of Q, increasing the distance when a trajectory has more sampled points. wDF would add points to the smaller trajectory using linear interpolation, what may create points that are not over the real movement.

In summary, several problems related to the heterogeneity of the data caused by the sampling process affect both precision and robustness of existing similarity measures.

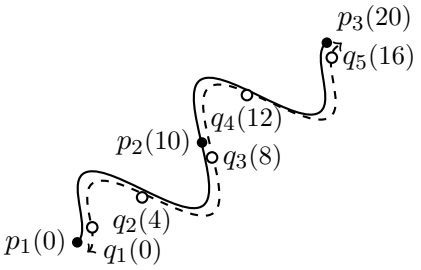


Figure 3. Trajectories P and Q with different sampling rates

3. Proposed Method - Foundations

In this section we explain the foundations of our proposal, that overcomes the aforementioned problems by using the time-geography ellipses proposed in (Pfoser and Jensen 1999), for each pair of sequential points for trajectory matching instead of building a radius of fixed size around every point, as detailed in Section 3.1. (Pfoser and Jensen 1999) use the maximum speed that an individual can move to define an upper bound for the ellipse, but we claim that individuals do not move at their maximum speed all time. As preliminary experiments have shown that the original time-geography ellipses tend to overestimate the ellipse sizes, and this overestimation significantly reduces the similarity accuracy, in Section 3.2 we propose a new distance function that approximates the upper bound of the movement to reduce the size of the ellipses in the trajectory representation.

3.1. *Creating Movement Ellipses*

Pfoser & Jensen in (Pfoser and Jensen 1999) proposed the use of ellipses to cover the area of *all possible locations of a moving object* between two consecutive sampled points, as shown in Figure 4(a). In that work the size of the ellipse is defined using the maximum velocity (v_{max}) that an individual can move until reaching the next sampled point. So the higher the time difference between two consecutive sampled points is, the bigger will be the size of the ellipse. Definition 3.1 shows how the ellipse is determined and an example is shown in Figure 4(b) (more details can be found in (Pfoser and Jensen 1999)).

Definition 3.1: Movement Ellipse. Given two points p_1 and p_2 they are set as the foci f_1 and f_2 of an ellipse e centered at c (we call f_1, f_2 and c the reference points of e). The eccentricity ε of the ellipse is the minimum distance between the foci, given by $\varepsilon = d_{euc}(p_1, p_2)$. The major axis μ_1 is the maximum distance that an object can travel at v_{max} during a period of time $\Delta_t = t_2 - t_1$, therefore $\mu_1 = v_{max} \times \Delta_t$. The thickness of the ellipse is given by the minor axis $\mu_2 = \sqrt{{\mu_1}^2 - \varepsilon^2}$.

An important remark is that in the extreme case, when the individual is traveling at v_{max} in a straight line ($v_{max} = v_{min} = \frac{\epsilon}{\Delta_t}$) the ellipse is degenerated to a line.

The main problem of the movement ellipse proposed by Pfoser and Jensen (1999) is that it assumes that v_{max} is known and that the individual could be traveling during the whole period at v_{max}, what is necessary to correctly obtain the upper bound of the movement. The problem of this assumption is that the maximum velocity of a moving object is valid only when it is traveling freely, but in most cases (e.g., cars) the trips are constrained by several factors such as: traffic flows, traffic lights, speed limits, road networks, road structure, climate condition, etc. Therefore, when working with real world trajectories it is impossible to define a single maximum velocity parameter that would

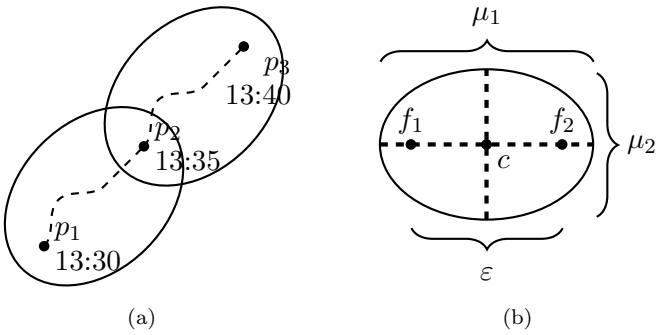


Figure 4. Example of: **(a)** Uncertainty representation using ellipses **(b)** Movement Ellipse with its foci (f_1 and f_2) and center (c) points, eccentricity (ε), major axis (μ_1), and minor axis (μ_2)

not overestimate the size of the ellipse. This overestimation is a problem for trajectory similarity. To overcome these limitations, in the next section we propose an approximate upper bound distance metric to dynamically define the size of the ellipses based on the intrinsic information of the trajectory.

It is important to highlight that we use the work proposed by Pfoser and Jensen (1999) solely as an inspiration to create dynamically-sized ellipses between two trajectory points instead of a fixed size radius around each trajectory point. The proposal of a new uncertainty model that covers all possible locations that an individual could visit between two sampled points is out of the scope of our work and, for completeness, we refer to the works of Kuijpers and Othman (2006), Trajcevski *et al.* (2010) and to Ranacher and Rousell (2013), that proposed an adaptive sampling approach based on the concept of time-geography ellipses for trajectory collection.

3.2. *Narrowing Movement Ellipses*

In this section we introduce the Approximate Upper Bound distance (AUB) for movement, a distance metric to estimate an approximate maximum possible movement length between two sampled points, which can be used to determine the size of the ellipse.

As stated before, in linear approximation the length of the movement between two sampled points is assumed to be equal to the shortest distance that an individual could travel between two points in space, i.e., the Euclidean distance, as illustrated with the distances between points (A, B) and (A, C) in Figure 5(a). Therefore, the use of the Euclidean distance is not appropriate to determine an upper bound distance between two sampled points.

Another possibility is to estimate an upper bound with the Manhattan distance, that is always greater than or equal to the Euclidean distance (Janssen 2007), but it tends to overestimate the travel distance between two points.

It is natural to assume that an approximate upper bound distance should be greater than the minimum possible movement length between two points where $v_{max} = v_{min}$ (in this case, the Euclidean distance). Therefore, with the exception of the situation when $d_{euc}(p_1, p_2) = 0$, in all other cases an approximate upper bound distance aub should be greater than the Euclidean distance ($aub(p_1, p_2) > d_{euc}(p_1, p_2)$). Even though in some cases the Manhattan distance (exemplified in Figure 5(b)) can be greater than the Euclidean distance, in other cases it can result in the same distance, i.e., equal to the minimum possible distance. It is easy to prove by counter-example that the Manhattan distance d_{man} does not always hold the property $d_{man}(p_1, p_2) > d_{euc}(p_1, p_2)$ with

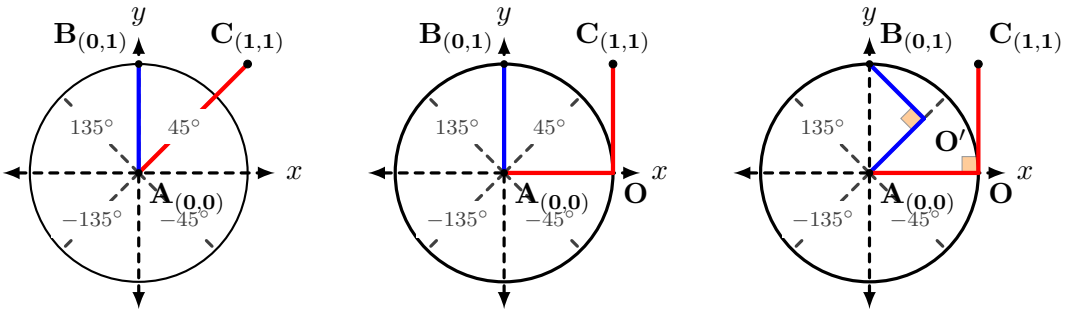


Figure 5. **(a) Euclidean Distance calculation for the pairs of points (A,B) (darker contiguous line) and (A,C) (lighter contiguous line):**
(A,B): $atan2(A,B) = 90°$; $d_{euc}(A,B) = |\overline{AB}| = 1$
(A,C): $atan2(A,C) = 45°$; $d_{euc}(A,C) = |\overline{AC}| = \sqrt{2}$
(b) Manhattan Distance calculation for the pairs of points (A,B) (darker contiguous line) and (A,C) (lighter contiguous line):
(A,B): $atan2(A,B) = 90°$; $d_{man}(A,B) = |\overline{AB}| = 1$
(A,C): $atan2(A,C) = 45°$; $d_{man}(A,C) = |\overline{AO}| + |\overline{OC}| = 2$
(c) Approximate Upper Bound Distance calculation for the pairs of points (A,B) (darker contiguous line) and (A,C) (lighter contiguous line):
(A,B): $atan2(A,B) = 90°$; $aub(A,B) = |\overline{AO'}| + |\overline{O'B}| = \sqrt{2}$
(A,C): $atan2(A,C) = 45°$; $aub(A,C) = |\overline{AO}| + |\overline{OC}| = 2$

$d_{euc}(p_1,p_2) > 0$, and in some cases it is equal to the Euclidean distance, as shown in Figures 5(a) and 5(b)), where both distances have the same value for the pair of points (A,B).

The Manhattan distance is adequate as an approximate upper bound when it reaches its maximum value in relation to the Euclidean distance, i.e., when the arc tangent between two points is equal to 45° (or to any of its corresponding values in the other quadrants of −45°, 135° and −135°), conforming an isosceles right-angled triangle ΔAOC as shown in Figure 5(b), where the hypotenuse $\overline{AC}$ is the Euclidean distance and the sum of the two other sides $\overline{AO}$ and $\overline{OC}$ is the Manhattan distance. However, the value of the Manhattan distance approximates the value of the Euclidean distance as the value of the arc tangent approximates 0 (in the interval $[0, 45[$) or 90 (in the interval $]45, 90]$) in the first quadrant (the same is valid for the corresponding angles in the other quadrants).

In order to avoid this approximation that follows the arc tangent variation between two points and tends to the equality between the Manhattan and Euclidean distance in its extremes (at 0° and 90°), we propose an Approximate Upper Bound distance (AUB) that gives a distance greater than the Euclidean distance in all cases (where $d_{euc}(p_1,p_2) > 0$) and greater than the Manhattan distance in all cases where $d_{man}(p_1,p_2) > 0$ and $atan2(p_1,p_2) \neq 45°$. For this distance, in all cases, an isosceles right-angled triangle is created from the straight line between two points, that is the hypotenuse, with the length given by the Euclidean distance). The length of the two other sides are the same and can be determined by using basic trigonometry ($|\overline{AO'}| = hypotenuse \times \sin 45°$) as illustrated in Figure 5(c), where $\overline{AB}$ and $\overline{AC}$ correspond to the hypotenuse, with length given by the euclidean distance, from which the 45° right-angled triangle is created. In Definition 3.2 we formally describe the Approximate Upper Bound Distance.

Definition 3.2: Approximate Upper Bound Distance. Given two points p_i and p_j, the line segment $\overline{p_i p_j}$ is assumed as the hypotenuse of an isocelles right-angled triangle.

The sum of the length of the other two sides of this triangle is the approximate upper bound distance between p_i and p_j. The formula is detailed in Equation 1.

$$\begin{aligned} aub(p_i, p_j) &= 2 \times \sin 45^\circ \times \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \\ &= 2 \times \frac{\sqrt{2}}{2} \times \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \\ &= \sqrt{2((x_i - x_j)^2 + (y_i - y_j)^2)} \end{aligned} \tag{1}$$

With the Approximate Upper Bound measure defined, we prove that it holds the aforementioned property of always being greater than the minimum distance between two points. Theorem 3.3 states that $aub(p_i, p_j) \geq d_{man}(p_i, p_j)$, as long as $p_i \neq p_j$. Also, Theorem 3.4 states that the approximate upper bound distance is always greater than the minimum possible distance $aub(p_i, p_j) \geq d_{euc}(p_i, p_j)$, considering that $p_i \neq p_j$.

Theorem 3.3: *Consider that the function $d_{man} : P \times P \to \mathrm{R}$ calculates the Manhattan distance. Then, for any pair of points $p_i, p_j \in P$, we have that $d_{man}(p_i, p_j) \leq aub(p_i, p_j)$.*

Proof: Let $a = |x_i - x_j| \geq 0$ and $b = |y_i - y_j| \geq 0$ and $p_i \neq p_j$. Then,

$$\begin{aligned} a + b &\leq \sqrt{2(a^2 + b^2)} \\ (a + b)^2 &\leq 2(a^2 + b^2) \\ a^2 + 2ab + b^2 &\leq a^2 + a^2 + b^2 + b^2 \\ 2ab &\leq a^2 + b^2 \end{aligned} \tag{2}$$

Expanding $0 \leq (a - b)^2$, we have $0 \leq a^2 - 2ab + b^2$, consequently $2ab \leq a^2 + b^2$ is implied. Therefore, $a + b \leq \sqrt{2(a^2 + b^2)}$. □ □

Theorem 3.4: *Consider that the function $d_{euc} : P \times P \to \mathrm{R}$ calculates the Euclidean distance. Then, for any pair of points $p_i, p_j \in P | p_i \neq p_j$, we have that $aub(p_i, p_j) > d_{euc}(p_i, p_j)$*

Proof: Let $a = |x_i - x_j| \geq 0$ and $b = |y_i - y_j| \geq 0$ and $p_i \neq p_j$. Then,

$$\begin{aligned} \sqrt{(a^2 + b^2)} &\leq \sqrt{2(a^2 + b^2)} \\ (a^2 + b^2) &\leq 2(a^2 + b^2) \end{aligned} \tag{3}$$

The proof is straightforward since two times any value greater than zero will always be greater than the value itself. □ □

3.3. *Creating Elliptical Trajectories*

In this section we redefine a movement ellipse considering the AUB distance to determine the size of the ellipse avoiding its superestimation, as described in Definition 3.5. An important remark is that we use ellipses solely with the intention of determining a dynamic matching threshold between sampled points, and do not use the ellipses to determine a movement model as in the original proposal of Pfoser and Jensen (1999).

Definition 3.5: Narrowed Movement Ellipse. Given two points p_1 and p_2 they are set as the foci of an ellipse e. The eccentricity ε of the ellipse is the euclidean distance between the foci, given by $\varepsilon = d_{euc}(p_1, p_2)$. The major axis μ_1 is given by the approximate upper bound distance between the points p_1 and p_2, therefore $\mu_1 = aub(p_1, p_2)$. The thickness of the ellipse is given by the minor axis $\mu_2 = \sqrt{{\mu_1}^2 - \varepsilon^2}$.

A sequence of narrowed movement ellipses conform an elliptical trajectory, as stated in Definition 3.6.

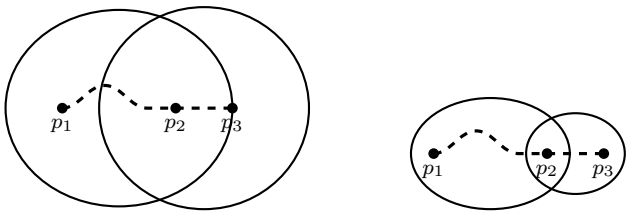
Definition 3.6: Elliptical Trajectory. An elliptical trajectory $E(T) \in \mathrm{E}$ is a time-ordered sequence of narrowed movement ellipses $\langle e_1, \ldots, e_n \rangle$ that belongs to a set of elliptical trajectories E.

Figure 6 illustrates a simple example that shows the effect of narrowing movement ellipses using the AUB distance. Let $p_1 = ((0,0),0), p_2 = ((10,0),1)$ and $p_3 = ((15,0),2)$ be three points of a trajectory $P = \langle p_1, p_2, p_3 \rangle$ representing the movement of an individual (in dashed lines). Assuming that the maximum speed the individual can reach is $20m/s$ we can determine the ellipses as follows:

- A movement ellipse e_1 (Figure 6(a)) is calculated as described in Definition 3.1: $\mu_1 = (v_{max} \times (t_2 - t_1)) = (20 \times 1) = 20$ and $\mu_2 = \sqrt{\mu_1^2 - \epsilon^2} = \sqrt{400 - 100} = 17.32$.
- A narrowed movement ellipse e'_1 (Figure 6(b)) is calculated as described in Definition 3.5: $\mu_1 = aub(p_1, p_2) = 14.14$ and $\mu_2 = \sqrt{\mu_1^2 - \epsilon^2} = \sqrt{199.93 - 100} = 9.99$.

The next ellipse e_2 and the narrowed ellipse e'_2 are analogously calculated. Notice in Figure 6 that both movement ellipses (e_1 and e_2) and narrowed movement ellipses (e'_1 and e'_2) cover the real movement, but the total area covered by the narrowed ellipses is $\approx 4\times$ smaller than the area covered by the movement ellipses.

The main benefits of using an elliptical representation of trajectories with narrowed ellipses is that they cover the real movement in the majority of the cases (as shown in the experiments of Section 5.1) and at the same time: i) are dynamically defined considering pairs of sequential points instead of using a radius with fixed user-defined size around each point; ii) avoid the underestimation of the movement as happens in linear approximation; and iii) avoid the overestimation of the ellipse area as happens when an individual is not traveling at its maximum speed.



(a) Movement Ellipses
$e_1 : \mu_1 = 20; \mu_2 = 17.32$
$e_2 : \mu_1 = 20; \mu_2 = 19.36$

(b) Narrowed Ellipses
$e'_1 : \mu_1 = 14.14; \mu_2 = 9.99$
$e'_2 : \mu_1 = 7.07; \mu_2 = 4.99$

Figure 6. From (a) Movement Ellipses to (b) Narrowed Movement Ellipses

4. Proposed Method - UMS

In this section we propose the UMS (Uncertain Movement Similarity) measure. Intuitively, the similarity score computed by UMS is based on three premises, where two elliptical trajectories are more similar if:

(1) Their shapes formed by the union of the movement ellipses look alike.
(2) Their movement ellipses cover the same space, sharing a greater common area.
(3) Their movement ellipses order represents individuals traveling in the same direction.

In order to cover these three premises we define the basis for the similarity score: *alikeness*, *shareness* and *continuity*.

The first concept is *alikeness*. Intuitively, the larger the area that the ellipses of two trajectories overlap is, the more alike they are. A simple way to obtain that would be through the evaluation of the intersection between the movement ellipses. However, it would not be accurate to only look if the ellipses intersect in cases where the trajectories start and/or end at different locations. An example is illustrated in Figure 7, where all the ellipses of S intersect at least one ellipse of R, but the shape of the trajectories is different. For this reason, instead of directly verifying if two ellipses intersect, we verify if the original trajectory points of one trajectory are spatially within the ellipses of the other, and vice-versa.

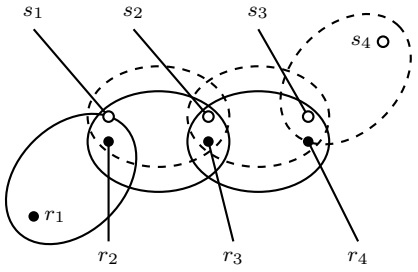


Figure 7. Example of Trajectories $R = \langle r_1, r_2, r_3, r_4 \rangle$ and $S = \langle s_1, s_2, s_3, r_4 \rangle$; and Elliptical Trajectories $E(R)$ (continuous ellipses) and $E(S)$ (dashed ellipses)

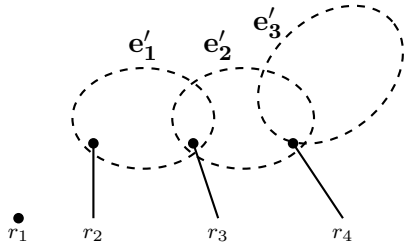


Figure 8. Example of a Trajectory $R = \langle r_1, \ldots, r_4 \rangle$ (black circles) and an Elliptical Trajectory $E(S) = \langle e'_1, e'_2, e'_3 \rangle$ (dashed ellipses)

Based on the trajectory R and the elliptical trajectory $E(S)$ we determine the point matching as follows:

Definition 4.1: Point Matching. Given a trajectory point $r \in R$ and an elliptical trajectory $E(S)$, r is said to match with $E(S)$ if it is within at least one of the ellipses $e' \in E(S)$. The result of the matching function $match(r, E(S)) : R \times \mathrm{E} \rightarrow \{0, 1\}$ (E is a set of elliptical trajectories) is computed according to Equation 4.

$$match(r, E(S)) = \begin{cases} 1 & \text{if } \exists e' \in E(S) | within(r, e') \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

For example, in Figure 8 we illustrate the trajectory R and the Elliptical Trajectory $E(S)$. In this case, it is easy to see that the point r_1 is not within any ellipse $e' \in E(S)$, hence it does not match with $E(S)$. Following this intuition, we formalize the concept of *alikeness* in Definition 4.2. Naturally, when more points $r \in R$ pair with ellipses $e' \in E(S)$, and more reference points $s \in S$ pair with ellipses $e \in E(R)$, the alikeness score will be higher.

Definition 4.2: Alikeness. Given two trajectories R and S the alikeness is computed by the function $\mathcal{A}(R, S) : \mathrm{T} \times \mathrm{T} \rightarrow [0, 1]$ (Equation 5).

$$\mathcal{A}(R,S) = \frac{\sum_{r \in R} match(r, E(S))}{length(R)} \times \frac{\sum_{s \in S} match(s, E(R)}{length(S)} \quad (5)$$

A high *alikeness* score is an indication that two elliptical trajectories have similar shapes in space. However, there are situations where two elliptical trajectories may have all its points matching the ellipses of the other trajectory but with less similar movements, as the example shown in Figure 9(b), where two individuals are traveling in parallel but different roads. The *alikeness* score considers only a binary value to quantify if the points of the trajectories match. However, binary values do not differentiate the proportion of intersection between two movement ellipses, as shown in Figure 9, where both $\mathcal{A}(R,S)$ and $\mathcal{A}(R,S)$ are equal to 1.

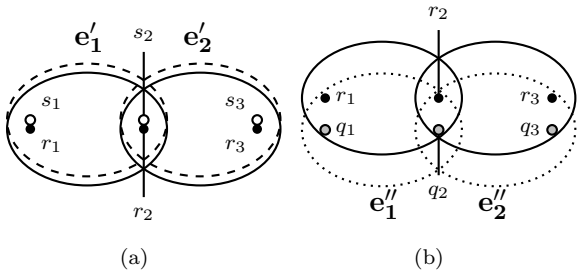


Figure 9. Example of trajectories R, S and Q, and their Elliptical Trajectories $E(R)$ (continuous ellipses), $E(S)$ (dashed ellipses) and $E(Q)$ (dotted ellipses). In this case both alikeness scores $\mathcal{A}(R,S)$ and $\mathcal{A}(R,Q)$ have the maximum value of 1.

A simple approach to solve this problem is to quantify the intersection area between the ellipses and choose an arbitrary value (e.g., the area of the bigger ellipse) to obtain a normalized score. However, the exact quantification of the intersection area of two rotated ellipses is very costly. For this reason, we define a function to quantify an approximate proportion of the intersection between two ellipses in Definition 4.4, that is faster and avoids the problem of binary distance assumption. According to this function, the nearest the points of a trajectory R and the reference points (foci and center points) of an elliptical trajectory $E(S)$ are (and vice-versa), the higher will be the sharing score.

Definition 4.3: Reference Point Normalized Distance. Given a point $r \in R$ and an ellipse $e' \in E(S)$, the reference point normalized distance d_{pnd} is given by the following Equation:

$$d_{pnd}(r, e') = \frac{\min_{k \in \{e'_{f_1}, e'_c, e'_{f_2}\}} d_{euc}(r,k)}{\xi(e')} \quad (6)$$

where $\xi(e')$ is the maximum possible distance for r to its nearest reference point of e' such that r is spatially within e'. This value is obtained from the equation $\xi(e') = \sqrt{\alpha^2 + \beta^2}$ where $\beta = \varepsilon/4$ and α can be obtained by finding half of the length of the chord parallel to the minor axis at a distance β using the equation $\alpha = \frac{1}{2}\mu_2\sqrt{1 - (\frac{\beta}{\mu_1/2})^2}$ derived from the ellipse general equation, as illustrated in Figure 10. By definition, if a reference point $r \in R$ is not within the ellipse e', we consider that $d_{pnd}(r, e') = 1$.

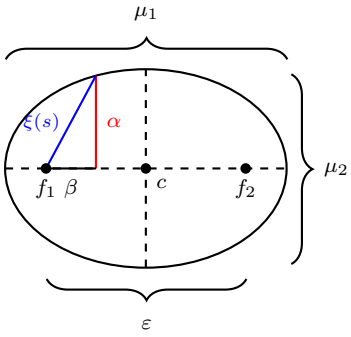


Figure 10. Example: calculating the value of ξ for an ellipse

Definition 4.4: Reference Point-Ellipse Sharing Score. Given a trajectory point $r \in R$ and an elliptical trajectory $E(S)$ the reference point-ellipse sharing score of r to $E(S)$ is the inverse of the minimum normalized distance between r and any ellipse $e' \in E(S)$, given by the following Equation:

$$share(r, E(S)) = 1 - \min_{e' \in E(S)} d_{pnd}(r, e') \tag{7}$$

For example, in Figure 9, the trajectory point $r_1 = (-25, 0)$ is only within the ellipse e'_1 that has the foci $f_1 = (-25, 2)$ and $f_2 = (0, 2)$. Therefore, to obtain the value of $share(r_1, E(S))$ we only need to calculate the value of the reference point normalized distance $d_{pnd}(r_1, e'_1)$. The value $d_{pnd}(r_1, e'_1)$ is the minimum distance from r_1 to the reference points $\{e'_{f_1}, e'_c, e'_{f_2}\}$ of e' normalized by the value $\xi(e')$. It can be computed according to the following steps: i) calculate the euclidean distance between r_1 and $\{e'_{f_1}, e'_c, e'_{f_2}\}$ to obtain the minimum distance. It is clear that the nearest point to r_1 is s_1, i.e., the foci f_1 of e', hence the euclidean distance $d_{euc}(r_1, e'_{f_1}) = 2$; ii) considering the foci of e' and the equations to compute the narrowed movement ellipse in Definition 3.5, we have that $\epsilon = 25$, $\mu_1 = 35.35$ and $\mu_2 = 25$. From these values it is possible to obtain $\beta = \epsilon/4 = 6.25$, $\alpha = \frac{1}{2}25\sqrt{1 - (\frac{6.25}{35.35/2})^2} = 11.7$; iii) from these values it is possible to compute the value of $\xi(e') = \sqrt{11.7^2 + 6.25^2} = 13.25$ by applying Pythagoras Theorem; and iv) we can compute the value of $share(r_1, E(S))$ as $1 - \frac{2}{13.25} = 0.85$. Similarly, the score $share(r_1, E(Q))$ can be calculated as 0.47, resulting in a lower score, as expected.

In Definition 4.4, we assume that the lower the normalized distance to the reference points is, the higher will be the proportion of the intersection between the ellipses of different trajectories, and consequently, the more similar they are. In Definition 4.5, we extend this assumption for the whole elliptical trajectory, and propose the concept of *shareness*.

Definition 4.5: Shareness. Given two trajectories R and S, the shareness is computed by the function $\mathcal{S}(R, S) : \mathrm{T} \times \mathrm{T} \rightarrow [0, 1]$ (Equation 8).

$$\mathcal{S}(R, S) = \frac{1}{2}\left(\frac{\sum_{r \in R} share(r, E(S))}{length(R)} + \frac{\sum_{s \in S} share(s, E(R))}{length(S))}\right) \tag{8}$$

The concepts of *alikeness* and *shareness* are useful to determine if two trajectories are

similar in space. However, none of these concepts take into consideration the order of the matchings. For instance, two individuals traveling in the same street but in opposite directions will have two elliptical trajectories with high *alikeness* and *shareness* score. The movement is not similar because the first point of one trajectory will match with the last ellipse of the other trajectory, the second point will match the penultimate, and so on for the following ellipses. In order to avoid this situation, we consider that beyond a high *alikeness* and *shareness* score, the matching order should be continuous for two trajectories to be considered similar.

To evaluate if the matchings of two trajectories are continuous, initially, we look for all the matchings of the points $r \in R$ with the elliptical trajectory $E(S)$, as stated in Definition 4.6.

Definition 4.6: Matching Sequence. Given a trajectory point $r \in R$ and an elliptical trajectory $E(S)$, the function $matchingSeq(r, E(S)) : R \times \mathrm{E} \rightarrow \mathcal{L}$ returns a list $\mathcal{L}$ with the positions of all elements $e' \in E(S)$ such that $within(r, e')$ is $true$, ordered according to the sequential index of matched ellipses.

A trajectory point can have several matchings with ellipses of another trajectory. For this reason, we use the function presented in Definition 4.7 to find the position of the first match of an ellipse e'_k that is greater than or equal to the last matching position of the previous ellipse e'_{k-1}. For example, if r_1 has the matching sequence $\langle 3, 4, 5\rangle$, and r_2 has the matching sequence $\langle 2, 4, 5, 6\rangle$, we have that $first(r_1, S) = 3$ and $first(r_2, S) = 4$.

Definition 4.7: First Matching Position. Given a point r_k and an elliptical trajectory $E(S)$, the position of the first matching of r_k is given by the function $first(r_k, E(S)) : R \times \mathrm{E} \rightarrow [-1, |E(S)|]$, according to the following conditions:

(1) if $k = 1 \ \wedge \exists s_l \in matchingSeq(r_k, E(S))$, the lowest value of l is returned;
(2) if $k > 1 \ \wedge \exists s_l \in matchingSeq(r_k, E(S))$, the lowest value of l such that $l >= first(r_{k-1}, E(S))$ is returned;
(3) otherwise, by definition, the value -1 is returned;

The first matching position of each point is considered in the definition of *continuity*. According to this concept, a sequence of first matching positions is created and verified for each trajectory. If more matching positions are ordered, that means that the matchings occurred in sequence, and the continuity score will be higher, as shown in Definition 4.8.

Definition 4.8: Continuity. Given two trajectories R and S, let $U = \langle first(r_1, E(S)), \ldots, first(r_{n-1}, E(S)),\ first(r_n, E(S))\rangle$ and $V = \langle first(s_1, E(R)), \ldots, first(s_{m-1}, E(R)), first(s_m, E(R))\rangle$ be two sequences with the first matching positions of all elements $r \in R$ and $s \in S$. Then, the continuity is computed by the function $\mathcal{C}(R, S) : \mathrm{T} \times \mathrm{T} \rightarrow [0, 1]$ (Equation 9).

$$\mathcal{C}(R, S) = \frac{\sum\limits_{0<i\leq|U|} valid(u_i)}{length(R)} \times \frac{\sum\limits_{0<j\leq|V|} valid(v_j)}{length(S)} \tag{9}$$

where $valid(u_k) : \mathrm{U} \rightarrow \{0, 1\}$ (analogously for $valid(v_k) : \mathrm{V} \rightarrow \{0, 1\}$) is the following function that states if a matching position is ordered w.r.t. the previous matching

position:

$$valid(u_k) = \begin{cases} 1 & \text{if } (k = 1 \wedge u_k \neq -1) \\ & \quad \vee (k > 1 \wedge u_k \neq -1 \wedge u_k >= u_{k-1}) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Finally, the similarity score of UMS is given by the composition of the three characteristics: *alikeness*, *shareness* and *continuity*, as shown in Definition 4.9. By definition if $length(E(R)) < 2$ or $length(E(S)) < 2$ we consider that $UMS(R, S) = 0$.

Definition 4.9: UMS Similarity. Given two trajectories R and S, the similarity score is computed by the function $UMS(R, S) : \mathrm{T} \times \mathrm{T} \rightarrow [0, 1]$ (Equation 11).

$$UMS(R, S) = \frac{(\mathcal{A}(R, S) + \mathcal{S}(R, S))}{2} \times \mathcal{C}(R, S) \quad (11)$$

UMS holds the properties of *non-negativity*, *reflexivity*, and *symmetry* (Lemmas 4.10 to 4.12).

Lemma 4.10: *(non-negativity). Given any two elliptical trajectories R and S, then $UMS(R, S) \geq 0$.*

Proof: Direct from Equations 5, 8, 9 and 11. □ □

Lemma 4.11: *(reflexivity). Given any two trajectories R and S, if $R = S$ and $length(E(R)) \geq 2$, then $UMS(R, S) = 1$.*

Proof: Definition 4.9 states that the similarity score is the result of the average matching and sharing scores multiplied by the continuity score. In all cases, if the elliptical trajectories have the same ellipses the score will be one. It happens because: i) all the trajectory points will intersect the ellipses of the other trajectory (alikeness score equal to 1); ii) at least one point will be at a distance 0 from a point of the other trajectory (shareness score equal to 1); and iii) the matchings will be ordered (continuity score equal to 1). It means that if two trajectories are equal ($R = S$) then all the terms of Equation 11 will reach the maximum possible value. Hence, if $R = S$ then $UMS(R, S) = 1$. □ □

Lemma 4.12: *(symmetry). Given any two trajectories R and S, in all cases $UMS(R, S) = UMS(S, R)$.*

Proof: Direct from Equations 5, 8, 9 and 11. □ □

5. Experiments

We performed experiments with two well-known datasets: i) the *mobility dataset* from the CRAWDAD Project[1] with low-sampled GPS trajectories (in average one point recorded each 61.8s) of 536 taxi drivers collected in May 2008 in San Francisco, USA; and ii) the Geolife dataset with more heterogeneous data and higher-sampled GPS trajectories (in average one point recorded each 16.2s) with daily life trajectories of 182 individuals using different transportation means (walking, cycling, driving, etc.), mainly collected in Beijing, China, between April 2007 and August 2012.

[1] http://crawdad.org/epfl/mobility/

Table 1. Summary of Datasets

Dataset	Trajectories	Points	Avg. Sampl. Distance	Avg. Sampl. Time
CRAWDAD	1,000,596	11,219,955	499m	61.8s
Geolife	59,158	24,876,978	96m	16.2s

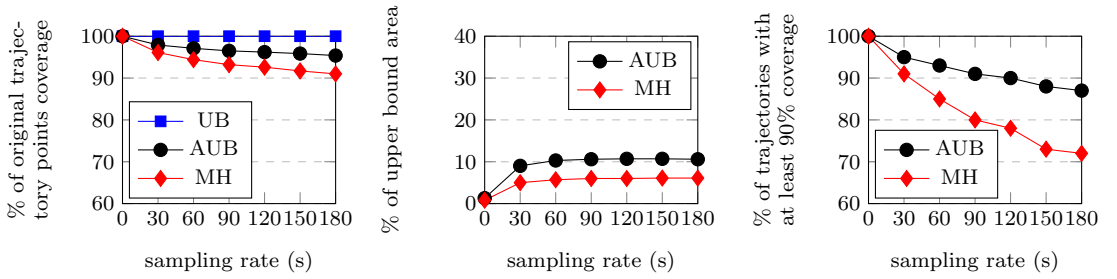
Originally, the CRAWDAD dataset contained one trajectory per driver (with duration of several days), what is not useful to determine similar movements around the town. For that reason, we applied two transformations to the dataset: i) trajectories were split each time the occupation status (taken or free) of the taxi changed; and ii) trajectories were split when a gap of 5 minutes between two consecutive points was found. As a result, the original 536 trajectories of the dataset were transformed in more than 1 million trajectories representing several taxi trips between different locations in San Francisco, USA. In the Geolife dataset we also applied the second transformation. In both datasets all the original points were kept. A summary of these datasets is presented in Table 1.

The experiments were performed comparing the approaches: DTW, LCSS, EDR, SWALE, wDF, CATS, EDwP and MSM. Different spatial threshold parameter values ($50m, 100m, 200m, 300m, 400m, 500m$) were used for the methods that require parameters, and the best results obtained for each method were reported. All these methods were implemented in Java and the experiments were performed with a i5 3317U processor, 8Gb of RAM and a solid-state disk. The results are shown in Sections 5.2-5.4.

Before we start the evaluation of the similarity measure, in Section 5.1 we evaluate the proposed Approximate Upper Bound distance function.

5.1. *Approximate Upper Bound Evaluation*

In Section 4.2 we claimed that AUB is a good approximate upper bound measure to create ellipses, balancing the coverage of the real locations between two sampled points, while at the same time it narrows the area determined by the upper bound ellipse approach.



(a) Coverage of the original trajectory points by the generated ellipses

(b) Area of generated ellipses as a % of the Upper Bound Area

(c) % of trajectories with at least 90% of the points covered by the generated ellipses

Figure 11. Results for the Evaluation of the Approximate Upper Bound Distance Measure for different sampling rates

We validate AUB using the Geolife dataset, which is very heterogeneous, has several kinds of trajectories collected with different transportation means, and contains highly-sampled trajectories, what allows the creation of transformed trajectories with varying sampling rates. In this experiment, we: i) selected all high-sampled trajectories, with an average sampling lower than 2s, and with a maximum speed of 200km/h (speed obtained

from real trajectories in the dataset), so we could set the maximum speed of the Upper Bound approach as 200km/h; ii) resampled these trajectories using several sampling rates (30s,60s,...,180s); and iii) created elliptical trajectories for all the low-sampled versions of the trajectories. Then, the ellipses average coverage of the real (originally sampled) locations between two sampled points and the average area size were computed for all trajectories[1].

Intuitively, a measure is more adequate to determine an approximate upper bound when it covers more points of the original trajectories with the smallest area size. To evaluate the results, we compare the proposed Approximate Upper Bound (AUB) measure with the Upper Bound Ellipse (UB), the original time-geography ellipses of (Pfoser and Jensen 1999), and the Elliptical Trajectories created using the Manhattan Distance as the major axis of the ellipses (MH).

The results of the experiments, detailed in Figures 11(a), 11(b) and 11(c), show that AUB had the best results regarding the balance between the coverage of the original trajectory points (y axis in Figure 11(a)), the area size (as a percentage of the upper bound area) covered by the generated ellipses (y axis in Figure 11(b)), and the percentage of trajectories with at least 90% of the original points covered considering a varying sampling rate (x axis in both graphs).

In summary, as shown in Figures 11(a), 11(b) and 11(c), AUB had an average coverage of 97% with an average area of 9% of the UB area (with $v_{max} = 200km/h$) and in average covered at least 90% of the original points of 92.4% of the trajectories, outperforming the Manhattan Distance that had averages of 94.1%, 5.1%, and 83.2%, respectively. AUB had a coverage closer to the upper bound value, as shown in Figures 11(a) and 11(c), with a slight increase in the area when compared to the Manhattan Distance, as shown in Figure 11(b)), resulting in a better balance between movement coverage and the size of the uncertain area representation. The explanation is that the Manhattan distance approximates linear interpolation in some cases, reducing the ellipse size and consequently its coverage (as demonstrated in Section 4.2).

In the following sections we evaluate UMS according to different aspects such as the precision of the measure in retrieval tasks (Section 5.2), the robustness of the measures regarding the sampling rate variation (Section 5.3) and the scalability (Section 5.4).

5.2. *Retrieval-based Evaluation*

In this section we extend the classic retrieval-based approach *precision at recall* (described in (Baeza-Yates and Ribeiro-Neto 1999)) to evaluate the precision of trajectory similarity measures.

5.2.1. Trajectory Precision@Recall Description

Initially, (i) two spatial regions $\mathcal{R}_1$ and $\mathcal{R}_2$ are defined as the origin and destination, respectively. The next step is (ii) to find, in the entire dataset D, a set $\mathcal{T}$ of trajectories that travel from $\mathcal{R}_1$ to $\mathcal{R}_2$, since several different routes may connect $\mathcal{R}_1$ to $\mathcal{R}_2$; (iii) a crowded route is selected and all trajectories of $\mathcal{T}$ that performed the movement in this route (being very similar to each other) are considered as the groundtruth set $\mathcal{G}$ with the

[1]For example: let $T_1 = \langle p_1, \ldots, p_{15} \rangle$ be the original trajectory with points sampled at each 1 second. Then, let $T_5 = \langle p_1, p_5, p_{10}, p_{15} \rangle$ be the resampled trajectory with 5 seconds as the rate. Finally, let $E = \langle e_1, e_2, e_3 \rangle$ be the elliptical trajectory representing T_5. The coverage of each ellipse is given by the percentage of points in the original trajectory that are spatially covered by the ellipse (e.g., the ellipse e_1 should spatially cover the points p_1, p_2, p_3, p_4 and p_5 of trajectory T_1) and the area size of the ellipses is calculated using basic geometry.

relevant trajectories. Next, (iv) the similarity score is computed between all the relevant trajectories in $\mathcal{G}$ with all trajectories of the entire dataset D, and (v) the trajectories are ranked by the similarity score. Naturally, a "perfect" measure would have at the top of the ranking all the trajectories in $\mathcal{G}$. Then, (vi) the precision is calculated for $size(\mathcal{G})$ levels of recall[1] and the average precision at each level of recall of all trajectories in $\mathcal{G}$ is calculated. Finally, (vii) the result is shown over a classical precision versus recall chart (Baeza-Yates and Ribeiro-Neto 1999).

Intuitively, the precision at a certain recall level is higher if more trajectories ranked until that recall level belong to $\mathcal{G}$ (consequently, a "perfect" measure would have the precision 1 at recall level 1, when all the relevant trajectories were returned). Other two alternatives to simplify the result in a single value include: i) compute the *mean average precision* (MAP) by considering the average precision in all levels of recall; and ii) find the break-even point (BEP) - the point where the precision and recall curve cross. In both cases, higher values indicate a better result of the measure.

5.2.2. Ground Truth Definition

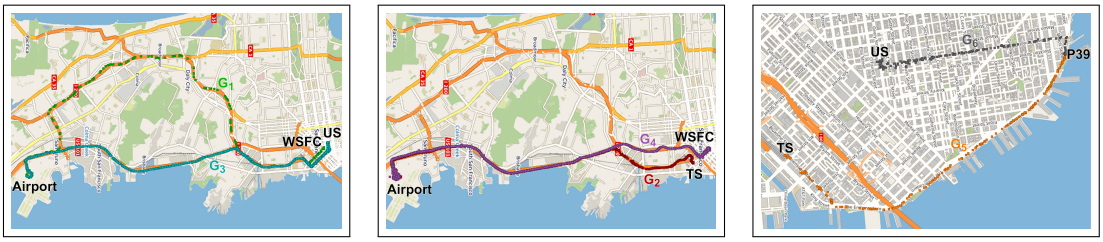
Initially, we selected five locations in San Francisco with a high density of pick-up/drop-off points near crowded places (Airport, Union Square (US), Train Station (TS), Pier 39 (P39) and Westfield San Francisco Center (WSFC)) in order to maximize the number of trajectories traveling between these regions in the CRAWDAD dataset. Then, we selected six of the most crowded routes with travels between these places, that are described in Table 2 as $\mathcal{G}_1$ to $\mathcal{G}_6$. The trajectories of the ground truth are illustrated in Figure 12. The use of a dataset with taxi trajectories was ideal to this kind of experiment because there are several trajectories between the pairs of regions that allow the construction of a ground truth with the trajectories in the same route, and at the same time there are several trajectories in slightly different routes, what makes the task of correctly ranking the most similar trajectories more challenging.

In the selection of the ground truth we considered several criteria to guarantee the variability in the evaluation including: i) a minimal number of trajectories in the route; ii) the average number of points of the trajectories; iii) the variation in the number of points (high in $\mathcal{G}_1$ and $\mathcal{G}_2$, medium in $\mathcal{G}_3$ and low in $\mathcal{G}_4$, $\mathcal{G}_5$ and $\mathcal{G}_6$); iv) the average sampling distance (high in $\mathcal{G}_1$ and $\mathcal{G}_4$, medium in $\mathcal{G}_2$ and $\mathcal{G}_3$, and low in $\mathcal{G}_5$ and $\mathcal{G}_6$); and v) characteristics of the route (the route taken in $\mathcal{G}_2$ was almost entirely in highways), while $\mathcal{G}_1$, $\mathcal{G}_3$ and $\mathcal{G}_4$ were partially inside the city and in $\mathcal{G}_5$ and $\mathcal{G}_6$ were entirely in roads inside the city).

Table 2. Description of Ground Truth Trajectories (Average reported with the Standard Deviation in the format $AVG \pm SD$)

	From	To	Trajectories	Avg. Points per Trajectory	Avg. Sampling Distance	Avg. Sampling Time
$\mathcal{G}_1$	WSFC	Airport	8	27±7.5	1536m±580	63.8s±22.3
$\mathcal{G}_2$	Airport	TS	90	26.4±11.7	1160m±383	59s±6.6
$\mathcal{G}_3$	Airport	US	99	22.7±4.3	1308m±186	56.5s±6.1
$\mathcal{G}_4$	WSFC	Airport	353	18.6±4.8	1590m±374	61.3s±11.9
$\mathcal{G}_5$	TS	P39	19	12.9±2.5	502m±102	61.5s±6.2
$\mathcal{G}_6$	P39	US	13	13.5±3	272m±57	64.8s±8.2

[1]For example, if $size(\mathcal{G}) = 10$ then the recall levels are $\{0.1, 0.2, \ldots, 0.9, 1.0\}$.



(a) (b) (c)

Figure 12. Groundtruth trajectories in: **(a)** $\mathcal{G}_1$: between the WSFC and Airport, and $\mathcal{G}_3$: between the Airport and Union Square (US); **(b)** $\mathcal{G}_2$: between the Airport and the Train Station (TS), and $\mathcal{G}_4$: between the WSFC and Airport; and **(c)** $\mathcal{G}_5$: between the Train Station (TS) and Pier 39 (P39), and $\mathcal{G}_6$: between the Pier 39 and Union Square

Table 3. Mean Average Precision (MAP) and Break-Even Point (BEP) for all the methods over each selected ground truth

	$\mathcal{G}_1$		$\mathcal{G}_2$		$\mathcal{G}_3$		$\mathcal{G}_4$		$\mathcal{G}_5$		$\mathcal{G}_6$	
	MAP	BEP	MAP	BEP	MAP	BEP	MAP	BEP	MAP	BEP	MAP	BEP
UMS	**0.66**	**0.63**	**0.41**	**0.43**	**0.64**	**0.63**	**0.49**	**0.50**	**0.72**	0.67	**0.40**	**0.40**
DTW	0.44	0.45	0.25	0.29	0.12	0.17	0.13	0.17	0.66	0.63	0.21	0.23
LCSS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EDR	0.34	0.43	0.07	0.14	0.07	0.12	0.08	0.15	0.43	0.43	0.17	0.21
Swale	0.34	0.45	0.09	0.15	0.07	0.13	0.08	0.15	0.43	0.47	0.18	0.24
wDF	0.41	0.43	0.09	0.13	0.07	0.13	0.07	0.12	0.53	0.52	0.16	0.21
CATS	0.27	0.35	0.10	0.14	0.07	0.11	0.04	0.08	0.24	0.31	0.18	0.21
EDwP	0.43	0.44	0.11	0.16	0.08	0.11	0.11	0.14	0.72	**0.68**	0.22	0.26
MSM	0.22	0.31	0.06	0.14	0.05	0.10	0.04	0.09	0.24	0.28	0.14	0.18

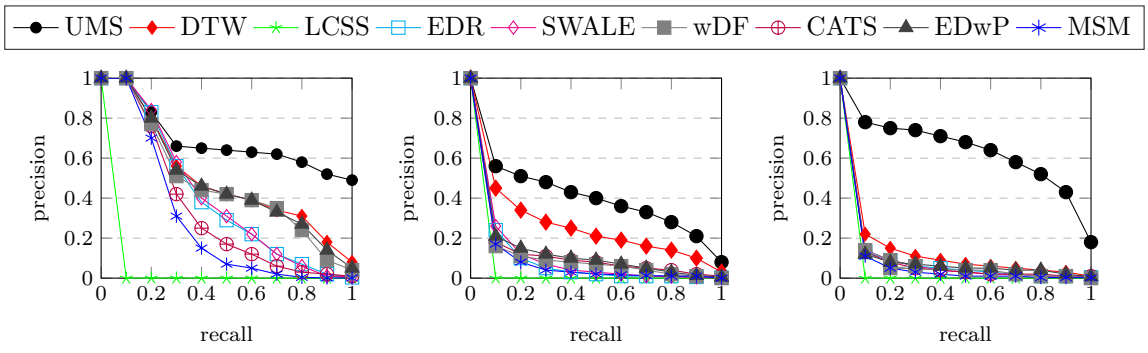
5.2.3. Retrieval-based Results

The $|\mathcal{G}_k|$ trajectories were used as the ground truth trajectories, and for each trajectory that belongs to $\mathcal{G}_k$, the $|\mathcal{G}_k|$ most similar trajectories should also belong to $\mathcal{G}_k$. For each one, a similarity search over the whole database was performed, ranking the trajectories until all $|\mathcal{G}_k|$ trajectories were found. The best result for a similarity measure is to return all trajectories in the groundtruth ranked in the positions from 1 to $|\mathcal{G}_k|$. The results of precision at each recall level is the average obtained for all $|\mathcal{G}_k|$ trajectories at that level. Figure 13 reports the results, and Table 3 summarizes the mean average precision (MAP) and break-even point (BEP) results for DTW, LCSS, EDR, SWALE, wDF, CATS, MSM, EDwP and UMS.

The results clearly show that UMS outperforms all the other methods in this dataset. The mean of the MAP and BEP results (MAP/BEP) of the six scenarios for UMS were 0.55/0.54, significantly higher than DTW (0.30/0.32), LCSS (0.00/0.00), EDR (0.19/0.25), SWALE (0.20/0.26), wDF (0.22/0.26), CATS (0.15/0.20), EDwP (0.27/0.29) and MSM (0.12/0.18).

The explanation of the worse results of related approaches relies on their limitations when dealing with real-world trajectory data. These approaches are very sensitive when dealing with low sampled trajectories (the average sampling rate in CRAWDAD dataset is 61.8 *seconds*), especially when the average sampling distance increases, what was the case in $\mathcal{G}_1$ to $\mathcal{G}_4$. It happens because they: i) use a fixed threshold around sampled points (LCSS, EDR, CATS and MSM); ii) directly sum the distance for pairs of points in different trajectories (DTW and wDF); or iii) try to interpolate points determining a rigid path in the missing parts of the trajectory (EDwP); while UMS uses ellipses with dynamically defined sizes to represent the movement between the sampled points, guaranteeing

a tolerance in the uncertainty representation and overcoming several problems pointed out in Section 2. The sensibility to these problems directly affects the similarity results in real-world datasets. For the fixed threshold, if its value is set too high (e.g., $500m$) the methods will match points of trajectories in other paths, while if the threshold is set too small (e.g., $50m$) it is not enough to match points of trajectories in the same route. The same happens with the measures that directly sum the distances, where the points of different trajectories in the same route can be sampled at different locations, increasing their distance. In the case of LCSS, the results are explained by the existence of very small trajectories that receive the maximum similarity score due to the lack of penalization for unmatched points. In addition, the use of interpolation presents some limitations, especially in low-sampled trajectories, where map-matching is not accurate (Zheng *et al.* 2012) and linear interpolation is too rigid, as adopted in (Ranu *et al.* 2015).



(a) 11-pt Precision @ Recall Curve for $\mathcal{G}_1$ (b) 11-pt Precision @ Recall Curve for $\mathcal{G}_2$ (c) 11-pt Precision @ Recall Curve for $\mathcal{G}_3$

(d) 11-pt Precision @ Recall Curve for $\mathcal{G}_4$ (e) 11-pt Precision @ Recall Curve for $\mathcal{G}_5$ (f) 11-pt Precision @ Recall Curve for $\mathcal{G}_6$

Figure 13. Detailed precision at recall results for the ground truths $\mathcal{G}_1$ to $\mathcal{G}_6$

For the above reasons, the precision difference between UMS and related approaches was higher in scenarios where the uncertainty is greater ($\mathcal{G}_1$ to $\mathcal{G}_4$), what demonstrates the effectiveness of the dynamic-sized ellipses approach to cover trajectory uncertainty. Although UMS also performs well in scenarios with less uncertainty ($\mathcal{G}_5$ and $\mathcal{G}_6$), while the difference in precision was smaller, it was also expressive, with the exception of DTW and EDwP in $\mathcal{G}_5$ that had similar results to UMS. It happened because at the same time the average sampling distance was smaller ($502m$), the standard deviation in the number of points was low, and most part of the trajectory was in an avenue parallel to the coast, with less options for trajectories in near parallel routes (as can be seen in Figure 12(c)). These results were not as good as in $\mathcal{G}_6$ because even though the route was straight between the two regions and the average sampling distance was small ($272m$) there were

several trajectories in different but parallel routes nearby that have wrongly received a score higher than the ground truth trajectories.

An important remark at this point is that some of the methods as LCSS, EDR, SWALE, CATS and MSM are not parameter-free, and that the difference between the best and worst score in several cases was significant. In addition, the best results had a parameter variation according to different ground truths and methods, what reinforces the fact that choosing the best parameters for different datasets with an unknown ground truth is not a trivial task.

In order to show the importance of AUB in the similarity accuracy of our measure UMS, we compare the similarity accuracy (precision) of datasets $\mathcal{G}_1$ to $\mathcal{G}_6$ using both the approximate ellipses (AUB) and the original time-geography ellipses proposed by Pfoser and Jensen (1999). The results for Pfoser and Jensen (1999) are $\mathcal{G}_1$: 0.20, $\mathcal{G}_2$: 0.05, $\mathcal{G}_3$: 0.05, $\mathcal{G}_4$: 0.09, $\mathcal{G}_5$: 0.65 and $\mathcal{G}_6$: 0.31. The results with AUB were: $\mathcal{G}_1$: 0.66, $\mathcal{G}_2$: 0.41, $\mathcal{G}_3$: 0.64, $\mathcal{G}_4$: 0.49, $\mathcal{G}_5$: 0.72 and $\mathcal{G}_6$: 0.40. Notice that the precision with AUB is much higher, because as demonstrated in Section 5.1, the use of maximum speed to determine the size of the ellipses tends to overestimate the covered area, and therefore resulting in large amounts of false positive ellipse intersection.

In summary, UMS had expressive higher mean average precision and break-even points than related approaches for all scenarios, with the difference to the other approaches being even more expressive in scenarios where the uncertainty was higher, confirming the effectiveness of ellipses on the coverage of the uncertain parts of trajectories.

5.3. *Robustness to Sampling Rate Evaluation*

In this section we evaluate the robustness of similarity measures in relation to the sampling rate. This experiment is similar to the one proposed in (Su *et al.* 2015) to validate the claim that existing similarity measures are very sensitive to sampling rate variation.

Initally, a set T with all the highly-sampled trajectories (sampling rate lower than 2s) of the Geolife dataset was selected. All trajectories were resampled using a variety of sampling rates ($5s, 15s, 30s, 45s, 60s, 90s$ and $120s$). Then, the set of trajectories resampled with $45s$ (T_{45}) was used as baseline and the distances of each trajectory in this set to their versions in the other sets ($T_5, T_{15}, T_{30}, T_{60}, T_{90}, T_{120}$) were calculated using DTW, LCSS, EDR, wDF, CATS, EDwP, MSM and UMS. The spatial threshold was set to $100m$ when required[1]. The resulting values of DTW, wDF and EDwP were normalized by the maximum distance value and SWALE was not included in the comparison because according to the values of reward and penalty it can have considerable variations (in the extremes having similar results to LCSS and EDR). A measure that is robust to variations in the sampling rate should have small distance variations.

The average distances between the trajectories of set T_{45} and its versions in the other sets are reported in Table 4, which is ordered by the average distance variation of all sets. UMS had the lowest variation, only 9%, while the second most robust measure (LCSS) had 16%. LCSS does not increase the distance when there are none-matching elements in the sequences and normalizes its score by the size of the smaller trajectory. It was able to match part of the points of the smaller trajectory with the $100m$ threshold, what reduces the score variation. All other methods had score variations greater than 25%.

In addition, the variation of the scores with LCSS was less uniform, varying between $0\% - 29\%$, while in UMS the difference varied between $0\% - 11\%$, as can be seen in the

[1]In the literature a radius of 100m or less is a common choice for these methods (Hung *et al.* 2015, Su *et al.* 2015).

boxplot score distribution shown in Figure 14. This experiment shows that the use of ellipses is less sensitive to sampling rate variations, once it covers the movement between two sampled points.

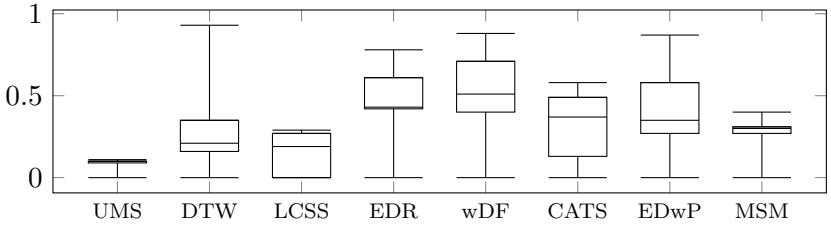


Figure 14. Sampling Rate Robustness: boxplot distribution for the similarity scores variation

Table 4. Average Distance to T_{45}

	T_5	T_{15}	T_{30}	T_{45}	T_{60}	T_{90}	T_{120}	**Avg.**
UMS	0.10	0.10	0.10	0	0.10	0.09	0.11	0.09
LCSS	0	0.12	0.23	0	0.29	0.19	0.27	0.16
MSM	0.30	0.27	0.27	0	0.31	0.30	0.40	0.26
DTW	0.93	0.35	0.18	0	0.16	0.21	0.29	0.30
CATS	0.13	0.28	0.37	0	0.47	0.49	0.58	0.33
EDwP	0.35	0.31	0.27	0	0.37	0.58	0.87	0.39
EDR	0.78	0.60	0.43	0	0.42	0.50	0.61	0.48
wDF	0.51	0.48	0.40	0	0.49	0.71	0.88	0.50

5.4. *Scalability Evaluation*

The time complexity of UMS is the same as DTW, LCSS, EDR, CATS, MSM and EDwP ($O(n * m)$). However, UMS performs little worse than some related approaches, because the use of ellipses instead of circles or the direct distance summing (that is based on the euclidean distance) requires more complex equations to determine if a point is within an ellipse (used to compute *alikeness* and *continuity*) and to compute the maximum distance within the ellipse to normalize the distances (used to compute *shareness*).

In order to evaluate the scalability, we perform the naive comparison of N randomly selected trajectories over the whole CRAWDAD dataset. Since the dataset has around 1 million trajectories, around $N \times 1M$ similarity computations were performed. All methods were implemented using dynamic programming. It is important to notice that, even though some methods have indexing techniques to reduce the number of complete comparisons in some scenarios, such as *top-k* queries, there are other scenarios, such as in clustering techniques, where the pairwise similarity computation of all trajectories is required. We also precomputed all the elliptical trajectories that were passed as parameter to the UMS method, what took around 17 seconds for the whole dataset.

Figure 15 shows the results in logarithmic scale. In average, the execution times for UMS were faster than EDwP and wDF, 1.1-1.2x slower than DTW, and 1.3-2.6x

slower than LCSS, EDR, MSM and CATS. However, it is important to highlight that the results of UMS are in the same order of magnitude of the other approaches and that its results showed greater precision in the retrieval-based experiments with a relatively small increase in the computation time, what shows a trade-off with considerable gains in precision with a little loss in performance. In addition, the last four methods require user-defined parameters as input, what adds the necessity of a parameter tuning step.

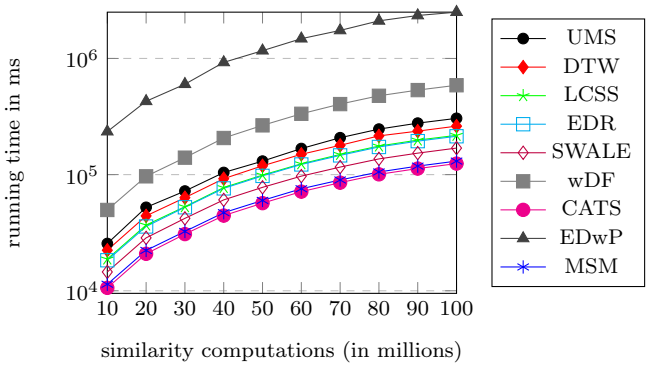


Figure 15. Scalability results: running time in miliseconds for a number of pairwise similarity computations (log10 scale)

6. Related Work

There are several methods developed for time-series similarity. The form of this kind of data (a sequence of elements) is similar to raw trajectories, what makes the use of these methods over raw trajectories straight-forward. **Dynamic Time Warping (DTW)** (Berndt and Clifford 1994) finds the best match between the elements of two sequences, creating a matrix with all possible combinations of two elements in the sequences and their distance as the entries. The total distance between the two sequences is the sum of the entries of the minimum contiguous path in the matrix. For this reason, DTW tends to be sensitive to noise, i.e., when a trajectory R has a point that is very distant from all the points of S, even if all the other points of R and S are close, the distance will be dominated by the noisy point.

Longest Common Subsequence (LCSS) (Vlachos *et al.* 2002) introduces a matching threshold when looking for the longest common subsequence. It reduces the effects of noise by quantifying the similarity between a pair of elements to binary values: 0 if the elements do not match and 1 otherwise. The *longest* matching sequence is used to calculate the similarity. A drawback of this approach is that it looks only the similar subsequence, ignoring possible gaps that may vary in size of the sequences.

Edit Distance on Real sequence (EDR) (Chen *et al.* 2005) is an evolution of LCSS, following a similar approach, where the distance between a pair of elements is quantized to binary values, and a matching threshold is used to avoid noise. EDR computes the distance of two sequences by adding 1 when the elements do not match and 0 when they match. Since this approach increases the distance for non-matching elements, it solves the problem of the gaps pointed as a drawback in LCSS, but as it matches points on a one-to-one basis, two similar trajectories with different number of points may have a high distance score.

Sequence Weighted Alignment (SWALE) (Morse and Patel 2007) presents a threshold matching approach that combines the idea of giving rewards to each match and penalties to matching gaps with user-defined parameters to determine the weights of each match/non-match. However, it does have the same limitation of EDR that considers matching points on a one-to-one basis and needs non-trivial user-defined parameters.

More recently, several other methods were proposed for raw trajectory similarity. **w-constrained discrete Frechet Distance (wDF)** (Ding *et al.* 2008) adapts the classical Frechet Distance (Alt and Godau 1995) to work with discrete series of points. This method adds temporal windows to the discrete Frechet Distance, in order to consider only the pairs of points that are within a given time-window. The distance between the trajectory points is directly calculated by a continuous distance function (e.g., euclidean distance). For that reason, as DTW, wDF is sensitive to noise. Another method based on the Frechet Distance was proposed by Buchin and Purves (2013), but instead of computing the Frechet Distance between the sampled points, it is computed over a set of space-time prisms generated over the sampled trajectory. The problem is that the space-time prisms are generated using the parameter v_{max} (maximal speed) to determine their sizes (as in the original work ofPfoser and Jensen (1999)), what in low-sampled trajectories tends to overestimate the size of the prisms, covering large areas and also affecting the accuracy of the similarity results.

Normalized Weighted Edit Distance (NWED) (Dodge *et al.* 2012) segments a trajectory in parts with homogeneous characteristics (e.g., same speed and/or direction), representing it as a sequence of symbols. This method considers two trajectories as similar if they have similar shapes or motion patterns, without considering the spatial dimension, what makes this approach useful only in scenarios where the geographic location of the movement is not relevant.

Clue-aware Trajectory Similarity (CATS) (Hung *et al.* 2015) proposes an approach that also considers matching thresholds (as LCSS and EDR), but instead of considering only binary values for each pair of points (match or not), it uses a function that considers the euclidean distance normalized over the spatial threshold value of the matched points. A limitation of this method is that the result is highly parameter dependent.

Edit Distance with Projections (EDwP) (Ranu *et al.* 2015) is the most related approach to our proposal. It is also a parameter-free method, that addresses the sampling variation problem by using a dynamic interpolation approach, which projects the points of the most dense trajectory in the interpolated lines of another trajectory to compute the scores based on a uniform sampled representation. A limitation of this approach is that it relies on the interpolation function, that in the absence of additional data usually is the classical linear interpolation as in (Ranu *et al.* 2015).

Multidimensional Similarity Measure (MSM) was proposed in (Furtado *et al.* 2016) for multidimensional sequences, including multidimensional trajectories. This measure considers a user-defined matching threshold for matching elements, similarly to EDR and LCSS, having the same limitations regarding the binary distance assumption. However, this approach is more flexible by allowing partial matchings and many-to-many element matching.

7. Conclusion

In this article we proposed a new trajectory parameter-free similarity measure, called UMS, which to the best of our knowledge, is the only approach that considers the interpolation error uncertainty of the movement beyond the use of linear interpolation when computing movement similarity, overcoming several limitations of the current methods when dealing with real trajectory data. Various evaluation techniques already proposed in the literature were used in the experiments and showed that UMS had better results regarding precision, accuracy, and robustness to sampling rate variations in the similarity assessment of uncertain trajectories when compared with state-of-the-art methods. In summary, the main contributions of this article are: i) a new distance function to estimate an approximate upper bound for movement uncertainty (AUB); ii) a new parameter free spatial similarity measure that covers the gaps between trajectory sampled points and; iii) the adaption of a classic retrieval-based evaluation technique to the movement data domain.

Acknowledgements

This work was partially supported by CAPES, CNPq and EU-IRSES-SEEK project (grant 295179). Nikos Pelekis and Yannis Theodoridis were partially supported by project datACRON, which has received funding from the European Union's Horizon 2020 research and innovation Programme (grant 687591).

References

Alt, H. and Godau, M., 1995. Computing the Fréchet distance between two polygonal curves.. *Int. J. Comput. Geometry Appl.*, 5, 75–91.

Baeza-Yates, R.A. and Ribeiro-Neto, B., 1999. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Berndt, D.J. and Clifford, J., 1994. Using Dynamic Time Warping to Find Patterns in Time Series.. *In*: U.M. Fayyad and R. Uthurusamy, eds. *KDD Workshop* AAAI Press, 359–370.

Buchin, M. and Purves, R.S., 2013. Computing similarity of coarse and irregular trajectories using space-time prisms. *In*: C.A. Knoblock, M. Schneider, P. Kröger, J. Krumm and P. Widmayer, eds. *21st SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2013, Orlando, FL, USA, November 5-8, 2013* ACM, 446–449.

Chen, L., Özsu, M.T., and Oria, V., 2005. Robust and Fast Similarity Search for Moving Object Trajectories. *In*: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, Baltimore, Maryland New York, NY, USA: ACM, 491–502.

Ding, H., Trajcevski, G., and Scheuermann, P., 2008. Efficient Similarity Join of Large Sets of Moving Object Trajectories. *In*: *Proceedings of the 2008 15th International Symposium on Temporal Representation and Reasoning*, TIME '08 Washington, DC, USA: IEEE Computer Society, 79–87.

Dodge, S., Laube, P., and Weibel, R., 2012. Movement similarity assessment using sym-

bolic representation of trajectories. *International Journal of Geographical Information Science*, 26 (9), 1563–1588.

Furtado, A.S., *et al.*, 2016. Multidimensional Similarity Measuring for Semantic Trajectories. *Transactions in GIS*, 20 (2), 280–298.

Hung, C.C., Peng, W.C., and Lee, W.C., 2015. Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. *The VLDB Journal*, 24 (2), 169–192.

Janssen, C., 2007. *Taxicab Geometry: Not the Shortest Ride Across Town.* Technical report, Iowa State University.

Kuijpers, B. and Othman, W., 2006. *In*: *Trajectory Databases: Data Models, Uncertainty and Complete Query Languages.*, 224–238 Berlin, Heidelberg: Springer Berlin Heidelberg.

Morse, M.D. and Patel, J.M., 2007. An Efficient and Accurate Method for Evaluating Time Series Similarity. *In*: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, Beijing, China New York, NY, USA: ACM, 569–580.

Pfoser, D. and Jensen, C.S., 1999. Capturing the Uncertainty of Moving-Object Representations. *In*: *Proceedings of the 6th International Symposium on Advances in Spatial Databases*, SSD '99 London, UK, UK: Springer-Verlag, 111–132.

Ranacher, P., *et al.*, 2016. Why GPS makes distances bigger than they are. *International Journal of Geographical Information Science*, 30 (2), 316–333 PMID: 27019610.

Ranacher, P. and Rousell, A., 2013. An Adaptive Sampling Approach for Trajectories Based on the Concept of Error Ellipses. *In*: *GI Forum 2013 Creating the GISociety – Conference Proceedings*, 169–176.

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 (3), 286–307 PMID: 27019646.

Ranu, S., *et al.*, 2015. Indexing and matching trajectories under inconsistent sampling rates. *In*: *2015 IEEE 31st International Conference on Data Engineering*, April., 999–1010.

Su, H., *et al.*, 2015. Calibrating Trajectory Data for Spatio-temporal Similarity Analysis. *The VLDB Journal*, 24 (1), 93–116.

Trajcevski, G., *et al.*, 2010. Uncertain Range Queries for Necklaces. *In*: *2010 Eleventh International Conference on Mobile Data Management*, May., 199–208.

Vlachos, M., Kollios, G., and Gunopulos, D., 2002. Discovering similar multidimensional trajectories. *In*: *Data Engineering, 2002. Proceedings. 18th International Conference on*, 673–684.

Zheng, K., *et al.*, 2012. Reducing Uncertainty of Low-Sampling-Rate Trajectories. *In*: *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ICDE '12 Washington, DC, USA: IEEE Computer Society, 1144–1155.