

Article

UTSM: A Trajectory Similarity Measure Considering Uncertainty Based on an Amended Ellipse Model

Ning Guo ^{1,2,*} , Shashi Shekhar ², Wei Xiong ¹, Luo Chen ¹  and Ning Jing ¹

¹ College of Electronic Science, National University of Defense Technology, Changsha 410073, China; xiongwei@nudt.edu.cn (W.X.); luochen@nudt.edu.cn (L.C.); ningjing@nudt.edu.cn (N.J.)

² Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, USA; shekhar@umn.edu

* Correspondence: guoning10@nudt.edu.cn

Received: 16 October 2019; Accepted: 13 November 2019; Published: 16 November 2019



Abstract: Measuring the similarity between a pair of trajectories is the basis of many spatiotemporal clustering methods and has wide applications in trajectory pattern mining. However, most measures of trajectory similarity in the literature are based on precise models that ignore the inherent uncertainty in trajectory data recorded by sensors. Traditional computing or mining approaches that assume the preciseness and exactness of trajectories therefore risk underperforming or returning incorrect results. To address the problem, we propose an amended ellipse model, which takes both interpolation error and positioning error into account by making use of the motion features of the trajectory to compute the ellipse's shape parameters. A specialized similarity measure method considering uncertainty called the Uncertain Trajectory Similarity Measure (UTSM) based on the model is also proposed. We validate the approach experimentally on both synthetic and real-world data and show that UTSM is not only more robust to noise and outliers, but also more tolerant of different sample frequencies and asynchronous sampling of trajectories.

Keywords: ellipse model; similarity measure; uncertainty; motion features

1. Introduction

Along with the massive new infrastructures being built for next-generation communication, GPS sensors are being installed by the millions every year in cars, cell phones, robots, and many other moving objects. These sensors are continuously recording carrier locations, generating a large number of trajectories of different types of objects. How to extract interesting information hidden in trajectory data is the focus of much research. Many advances have been made in trajectory modeling and pattern mining, including work on measuring trajectory similarity.

Trajectory similarity measures are fundamental for spatial movement clustering or classification [1,2], and similarity queries can be applied broadly in many spatial data mining applications, such as finding the group/flock pattern of moving objects, detecting spatiotemporal outliers, and predicting weather or climate such as a hurricane route. Information about the similarities in movement patterns can enable traffic managers to find suspicious behavior and increase safety and security [1].

However, real-life trajectory data are far from being reliable enough for many applications because datasets collected from mobile sensors are often imprecise and incorrect due to noise and discontinuous records [3]. In the real world, the dynamics of moving objects in their spatiotemporal reference frame are macroscopically continuous and accurate. During the process of recording, propagation, and calculation, however, these dynamics are expressed as discrete spatiotemporal sampling points or interpolated continuous polylines, which inevitably brings some uncertainty. In most cases, the needs of analysis and application can be met despite some uncertainty in the data. However, in some

scenarios, due to factors such as sensor failure or environmental influence, the sampling frequency does not meet the requirements of Nyquist's sampling law, which may cause large errors in the trajectory rebuilding and lead to erroneous computing results. Even if we could collect precise trajectory data, we cannot assume they will remain accessible due to the increasing privacy protection demands of consumers. Thus, there is also a need for new models and algorithms that work with blurred data.

Another major concern in measuring the similarity between trajectories is that different sampling frequencies and an asynchronous sampling strategy will cause misjudgments. Consider the similarity comparison shown in Figure 1: the real trajectories of P and Q in Figure 1a are closer than those of P and R. However, because of the low sampling frequency and heterogeneous distribution of the sampled points, the rebuilt trajectories in Figure 1b obtained by linear interpolation show the opposite result: R is more similar to P than Q.

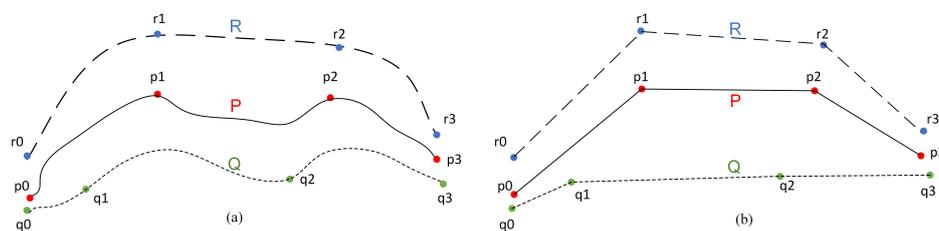


Figure 1. A trajectory similarity distortion example.

To reduce this kind of misjudgment under the current data quality condition, we must take uncertainty into consideration. In this work, our main contributions are outlined as follows:

- We propose an amended ellipse model derived from [4] that could better describe the uncertainty in a trajectory by dynamically computing the model's shape parameters based on the motion features of each segment of the trajectory.
- We present a new similarity measure based on the model, called the Uncertain Trajectory Similarity Measure (UTSM). Validated by experiments on both synthetic data and real-world data, UTSM shows a better ability to compare similarity between trajectories, and it is also more robust to noise and outliers and more tolerant of different sample frequencies and asynchronous sampling.

The remainder of the paper is organized as follows: Section 2 introduces basic concepts related to trajectory similarity metrics, gives the formal problem statement, and provides a table of the symbol conventions used in the paper; Section 3 reviews related work and current technology; our re-designed ellipse model and similarity measure method are proposed in Section 4; Section 5 presents the validation experiments and results, while Section 6 concludes the paper and looks ahead toward future work.

2. Preliminaries and Problem Definition

2.1. Basic Concepts

- Spatiotemporal Trajectory

A space-time trajectory is a continuous curve formed by the motion of a moving object in Euclidean space over a certain period. The morphology of the trajectory can be accurately described by a continuous function. However, in reality, the spatial position of the moving object is generally recorded by a sensor at a fixed or random frequency. The object's trajectory is usually represented by a sequence of positions containing spatial and temporal information, such as:

$$P \{(s_0, t_0), (s_1, t_1), \dots, (s_{n-1}, t_{n-1})\}$$

where n is the number of sampling points of the entire trajectory and s_i is the position of the moving object at time spot t_i . In practical applications, s_i is generally a 2D or 3D coordinate. This paper mainly focuses on trajectories in two-dimensional space, so combined with the time information, the sampling point can also be expressed as (x_i, y_i, t_i) .

- **Interpolation Error**

Since sensor recorded data are a sequence of sampling points that are not continuous in either time or space, the sampling points are generally connected in chronological order to describe the overall morphology of the trajectory. The data gap between adjacent sampling points is usually filled by linear interpolation, and the trajectory is transformed into a polygonal line. Interpolation is essentially an estimate and approximation of missing data and will inevitably introduce uncertainty in the gap between adjacent sampling points. The error introduced by this procedure is called Interpolation Error (InE) in this paper.

- **Positioning Error**

Due to the accuracy limit of the positioning sensors or different signal intensities of different areas, the positioning data of the sampling points in a trajectory generally have a certain amount of error. This is also called measurement error. The mean spatial accuracy of a standard GPS receiver is near two meters horizontally at a 95% confidence interval [5] and worse in occluded areas because of the weak signal. In this paper, this kind of error is defined as Positioning Error (PoE).

- **Trajectory Similarity**

A similarity measure or similarity function is generally a real valued function that quantifies the similarity between two objects. Trajectory similarity refers to the degree of similarity between a pair of trajectories, including their spatiotemporal position, shape trend, motion characteristics, etc., which together can measure the overall similarity of their movement. Distance is a typical similarity measure, such as Euclidean distance, Hausdorff distance, Frechet distance, etc. However, distance is inversely proportional to similarity. The smaller the distance between two trajectories, the greater their similarity. An artificially defined similarity model can also be used to express the similarity between trajectories. In practical applications, similarity metrics are often normalized to the interval of $[0, 1]$ for heterogeneous comparison.

The performance of different trajectory similarity measures will be further discussed in Section 3.

2.2. Problem Statement

Our problem is defined formally as follows: Given two trajectories P and Q , find a normalized similarity measure denoted as $S(P, Q)$ for the two trajectories, which meets the following requirements:

$$\begin{aligned} 0 < S(P, Q) &\leq 1 \\ S(P, Q) &= S(Q, P) \\ S(P, P) &= 1 \end{aligned} \tag{1}$$

In other words, the metric is non-negative, symmetric, and reflexive. In addition to the usual constraints, we also need to consider the two kinds of uncertainty, i.e., interpolation error and positioning error, when designing the measurement model. For simplicity, we assume that the time intervals between consecutive pairs of anchor points are equal.

The notations we use are listed in Table 1

Table 1. Notations and definitions.

Notation	Definition
V_m	Maximum velocity of a moving object
Δt	Time interval between two consecutive points
$d_{euc}(p, q)$	Euclidean distance between point p and point q
InE	Interpolation error
PoE	Positioning error
$E(p_i, p_{i+1}), E_i$	Elliptical uncertain area of a segment, which is also the projection on a 2D location plane of the corresponding bead body.
$E(P)$	The union area of all the ellipses obtained from every pair of consecutive points.
$S(P, Q)$	Similarity between trajectory P and Q
$NS(P, Q)$	Normalized similarity between trajectory P and Q

3. Related Work

3.1. Spatiotemporal Trajectory Data Models

Trajectory models can be said to date back to the 1970s when researchers were trying to find a more comprehensive and effective way to describe climate data and forecast climate and weather change. In 1972, a numerical three-dimensional trajectory model for wind data was proposed in [6], which was operational for the prediction of temperature and dew point. From then on, numerous trajectory data models have been proposed, such as some models derived from classic general spatiotemporal data models like STER [7], MADDS [8], and other featured models designed under particular application backgrounds like the constraint model [9], stop-move model [10], and event-based model [11]. In most instances, only the location information at timestamps is used and is assumed to be exact. Some research has focused on the semantic information of moving objects [12–14]. Among them, there is one particular model that deserves special attention called the multi-granularities model [15]. It uses a bead-like shape to describe the possible path area that a moving object may take between two anchor points. Two key concepts behind this model are the lifeline bead and the lifeline necklace. The lifeline bead model is derived from the time geography framework [16]. Its mathematical foundation was further elaborated in [17–22] including a spacetime path and a spacetime prism, two fundamental elements of time geography. The spatiotemporal bead model successfully accounts for the uncertainty introduced by the sampling and interpolating representation of trajectory data. Our proposed uncertainty model is theoretically based on the bead model.

3.2. Uncertainty in Spatiotemporal Trajectories

Most previous research addressing uncertainty in trajectories has focused on interpolation error [23,24]. Existing models and methods that consider spatial uncertainty include the circular range model [25], the cylinder model [26], the grid model [27], the bead model [15] (also called multi-granularities model) or the ST-prism model [17,19], etc., which mainly describe the possible positions of moving objects between adjacent sampling points and their distribution probabilities in different geometric forms. Beads are not easy to handle due to their 3D body, which combines both spatial and temporal metrics. A classical applied method is to project a bead onto a 2D-plane as an ellipse, as shown in Figure 2. Much research has been conducted based on the projected ellipse, such as accessibility computing and location distribution possibility prediction [28–30].

The usual practice for handling uncertainty in trajectories is to assume that the positions of the moving objects in the time intervals between the sampling points conform to a certain distribution. For example, one approach [31] first discretizes the time-space between the sampling points and then adopts a random-walk method to simulate the position of the moving object between two fixed sampling points, so that an approximate access probability distribution on the discrete spatiotemporal

grid can be obtained. Another work [28] improved the use of the random-walk method and the Brownian motion model and described the probability distribution of moving objects in a Potential Path Area (PPA) in a continuous spatiotemporal domain. However, both of them ignore the positioning error of anchor points [4].

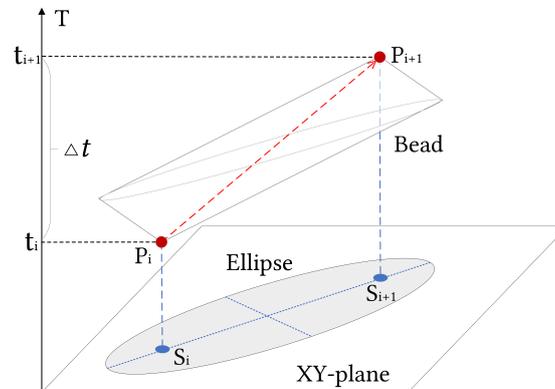


Figure 2. ST-bead and planar-ellipse.

3.3. Trajectory Similarity Measures

3.3.1. Geometric Distance Metric

Distance is a common metric of object similarity. It is generally assumed that distance is inversely proportional to similarity. Minkowski distance measures such as Euclidean distance, Manhattan distance, Chebyshev distance, and their variants are intuitive measures of trajectory similarity. There are also some other specifically designed measurement methods like cosine distance and Hausdorff distance, which focus on geometric features, or Hamming distance, Jaccard distance, and correlation distance, which focus on statistical properties. The area between a pair of trajectories can also be used as a distance measure [32]. However, most of these methods ignore the temporal property of trajectories.

3.3.2. Time Based Distance Metric

To overcome the shortage of geometric based metrics and take time or temporal order into account, numerous time based distance or similarity measures have been proposed, such as Frechet distance, DTW [33], LCSS [34], ERP [35], EDR [36], Swale [37], STLIP [38], and NWED [39]. Much research has been done to improve the computational efficiency of these classical methods or to adapt their measurement models for practical application. However, most of the solutions assume that a trajectory is exact in its location property.

3.3.3. Similarity Measures that Consider Uncertainty

Recent work [4] improved the original elliptical trajectory model [23] to describe the reachable range of a trajectory path. It eliminated the maximum speed assumption in the original model and replaced it with an approximate upper bound distance. It then defined and calculated measures of likeness, sharedness, and continuity. Finally, a comprehensive similarity expression was obtained. However, they simply expanded the Euclidean distance of the adjacent anchor points by a fixed coefficient to determine the parameters of the ellipse. In other words, all the ellipses had the same eccentricity, which cannot describe the real situation of uncertainty along the trajectory. The work of [40] proposed a method of quantifying trajectory similarity as an interval, rather than a single value, called trajectory interval distance estimation, to capture the uncertainty that results from different sampling rates and asynchronous sampling. The estimation model was based on a circular bounding area with a radius proportional to the estimated maximum speed. While this circular bounding area

could tackle the problem of asynchronous sampling to a certain degree, the maximum speed was arbitrarily determined to be the max value of the average speed of each pair of consecutive anchor points. There are also some fuzzy set similarity measures based on the fuzzy theory that have been applied in shape classification [41].

To sum up, related work about the similarity measure is shown as Figure 3.

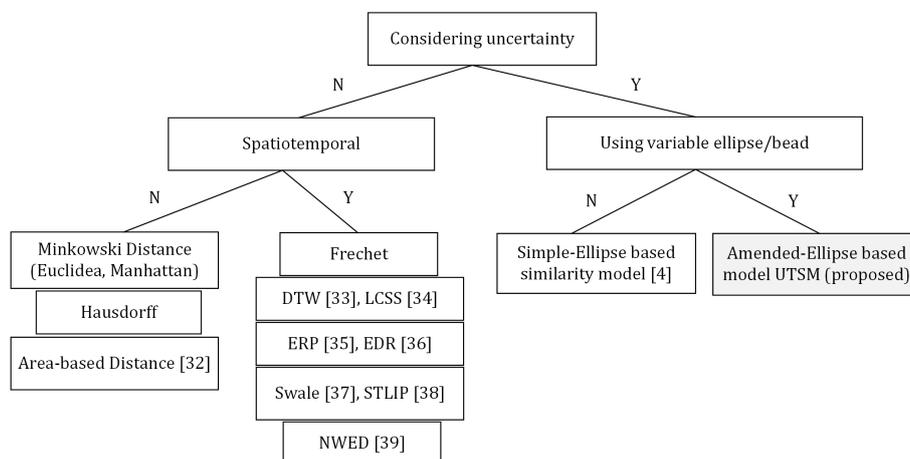


Figure 3. Trajectory similarity measure methods.

To overcome this limitation of previous work, we propose a novel amended ellipse model that represents a trajectory as a chain of interconnected ellipses, whose shape parameters are dynamically calculated by the motion features of adjacent segments. We then use the new model to design a similarity measure that is more robust to both interpolation error and position error.

4. Proposed Approach

4.1. Amended Ellipse Model

The traditional bead model considers the possible path that a moving object can take between two consecutive anchor points (also called sample points) [15]. The slope of the bead is determined by the maximum velocity of the moving object. However, in real cases, moving objects are not likely to keep moving at maximum speed. Therefore, in a 2D projection plane, the shape of the projected ellipse should not always be determined by the maximum velocity, which is often assigned arbitrarily. In our work, to improve the rationality of describing the possible position between anchor points, we take the movement features into account when determining the shape parameters of an uncertain ellipse.

4.1.1. Initial Model without Positioning Error

As the graph in Figure 4 shows, we take two adjacent anchor points as the focal points of $E(p_i, p_{i+1})$. What remains unknown of the ellipse is the length of the semi-minor axis, viz. b . Intuitively, b indicates the uncertainty of the vertical range of the moving object due to interpolation error. The core of our method is to estimate this range using motion features reflecting adjacent segments.

The interpolation error originates from the linear interpolation, the most commonly used interpolation method, which results in the failure of reflecting the change of movement states. Any direction change between two consecutive anchor points is missed and is simply replaced by a straight line. The range of a possible path area is related to the speed and the change in moving direction. We can use the related motion features of neighborhood segments to approximate the real situation. In addition to velocity and acceleration, there are other important features such as sinuosity and turning angle that can reflect motion characteristics, especially the movement direction and its alteration. Since the ellipse model is based on segments rather than single points, we use a synthesis method to compute the possible vertical range of the movement and estimate the value of b .

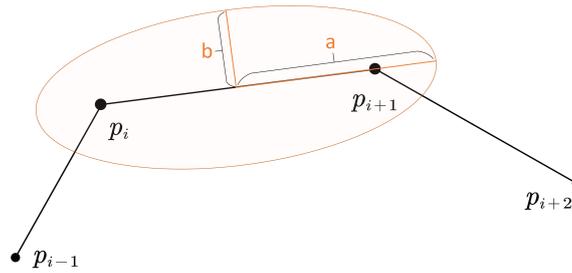


Figure 4. Uncertain ellipse of an interpolated segment.

Take $segment(p_i, p_{i+1})$ in Figure 5 as an example, the turning angle is derived from the direction difference between the average velocity (in terms of vectors) of $segment(p_{i-1}, p_{i+1})$, $segment(p_i, p_{i+1})$, and $segment(p_i, p_{i+2})$, i.e.:

$$\begin{aligned} \theta_i^1 &= |\angle(p_{i-1}, p_{i+1}) - \angle(p_i, p_{i+1})| \\ \theta_i^2 &= |\angle(p_i, p_{i+2}) - \angle(p_i, p_{i+1})| \end{aligned} \tag{2}$$

where \angle means the angle between the vector and the horizontal line.

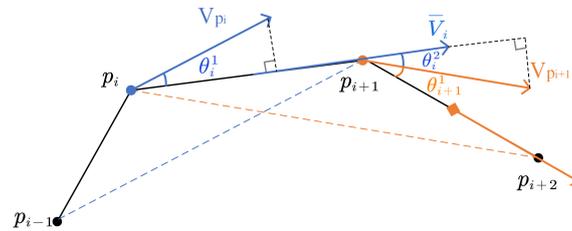


Figure 5. Motion features of a segment.

An intuitive hypothesis is that the vertical uncertainty range is proportional to the velocity component in the vertical direction. As Figure 6 shows, the vertical component of velocity $v_{p_i} \sin \theta_i^k$ can result in a vertical distance of $\frac{1}{2} v_{p_i} \sin \theta_i^k \cdot \frac{\Delta t}{2}$ until the component decelerates to zero. The deceleration time is set to $\frac{\Delta t}{2}$ because the moving object has to return to anchor point P_{i+1} . We can get two uncertain vertical distances derived from the starting velocity and ending velocity of the segment. Therefore, we can estimate the value of the semi-minor axis b by computing the average of the two uncertain vertical distances:

$$\begin{aligned} b &= \frac{1}{2} \left(\left| \frac{v_{p_i} \sin \theta_i^1}{2} \right| \cdot \frac{\Delta t}{2} + \left| \frac{v_{p_{i+1}} \sin \theta_i^2}{2} \right| \cdot \frac{\Delta t}{2} \right) \\ &= \frac{1}{8} \Delta t \left(v_{p_i} |\sin \theta_i^1| + v_{p_{i+1}} |\sin \theta_i^2| \right) \end{aligned} \tag{3}$$

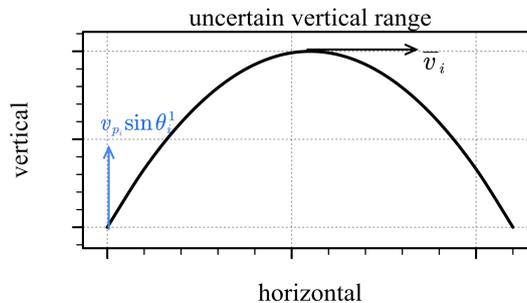


Figure 6. Uncertain vertical range estimation based on turning angle.

The amended ellipse parameters are thus:

$$\begin{aligned}
 c &= \frac{1}{2}d_{euc}(p_i, p_{i+1}) \\
 b &= \frac{1}{8}\Delta t \left(v_{p_i} |\sin \theta_i^1| + v_{p_{i+1}} |\sin \theta_i^2| \right) \\
 a &= \sqrt{b^2 + c^2}
 \end{aligned}
 \tag{4}$$

Examples:

- If a segment has roughly the same direction as both its neighbors, its b will be very small, and the eccentricity would be rather large, i.e., the segment’s area of uncertainty is a narrow ellipse (Figure 7), which is consistent with our usual perception that an object moving in a straight line tends to have smaller location uncertainty. For the similarity measure proposed in the following part, these “straight” kinds of curves tends to have less similarity with each other because of the smaller area of uncertainty.



Figure 7. A “straight” line with narrow ellipses.

- A segment with large turning angles to its neighbors will have a bigger b, contributing to a smaller eccentricity and a wider ellipse (Figure 8). The phenomenon is also consistent with the intuitive perception that an object moving at sharp angles tends to have greater location uncertainty. For the proposed similarity measure, these “winding” kind of curves have a greater tendency to have greater similarity with each other as a result of the bigger overlapping area of adjacent uncertainty extents.

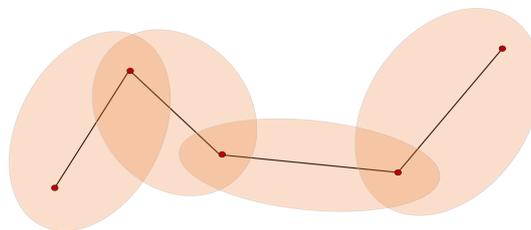


Figure 8. A “winding” line with wide ellipses.

4.1.2. Model Considering Positioning Error

Introducing positioning error will add an additional offset angle to the original turning angle of the segment, as shown in Figure 9.

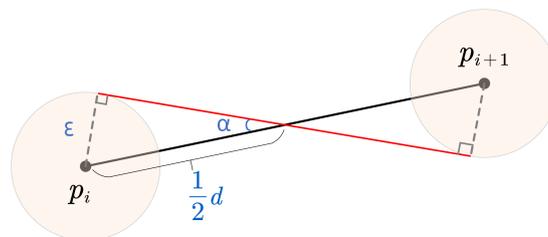


Figure 9. Offset angle caused by positioning error.

The offset angle caused by the positioning error can be obtained by:

$$\sin \alpha = \frac{\epsilon}{\frac{1}{2}d} = \frac{2\epsilon}{d}
 \tag{5}$$

We use the mean spatial accuracy of a standard GPS receiver as the position measurement error, i.e., $\varepsilon = 2m$ [5]. In most cases, the error is far less than the distance between two consecutive anchor points; $\frac{2\varepsilon}{d}$ is a very small value, and so is α . According to Taylor series expansion:

$$\sin \alpha = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} \alpha^{2k+1} = \alpha - \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!} - \dots \stackrel{\alpha \rightarrow 0}{\approx} \alpha \quad (6)$$

Therefore, under the circumstance of a regular position measurement error, we can approximately obtain: $\alpha \approx \sin \alpha = \frac{2\varepsilon}{d}$, and the amended turning angle of the segment is $\theta_i^k + \frac{2\varepsilon}{d}$, $k = 1, 2$.

Therefore, the final amended ellipse parameters that consider both InE and PoE are:

$$\begin{aligned} c &= \frac{1}{2} d_{euc}(p_i, p_{i+1}) = \frac{1}{2} d \\ b &= \frac{1}{8} \Delta t \left[v_{p_i} \sin \left(\theta_i^1 + \frac{2\varepsilon}{d} \right) + v_{p_{i+1}} \sin \left(\theta_i^2 + \frac{2\varepsilon}{d} \right) \right] \\ a &= \sqrt{b^2 + c^2} \end{aligned} \quad (7)$$

If the anchor points are extremely close to each other, we can use the original formula of offset angle $\alpha = \arcsin \frac{2\varepsilon}{d}$. Therefore, the semi-major axis b will be:

$$\begin{aligned} b &= \frac{1}{8} \Delta t \left[v_{p_i} \sin \left(\theta_i^1 + \arcsin \frac{2\varepsilon}{d} \right) + \right. \\ &\quad \left. v_{p_{i+1}} \sin \left(\theta_i^2 + \arcsin \frac{2\varepsilon}{d} \right) \right] \end{aligned} \quad (8)$$

4.2. Uncertain Trajectory Similarity Measure

According to our amended ellipse model, a trajectory can be represented by a chain of ellipses with different shape parameters. A similarity model is designed based on the uncertainty description. First, we need to define a key concept called match. If an anchor point p_i in P is located inside the ellipse of segment (q_j, q_{j+1}) in Q , then p_i is matched to (q_j, q_{j+1}) or p_i is matched to $E(Q)$, as Figure 10. Otherwise, p_i is mismatched to Q .

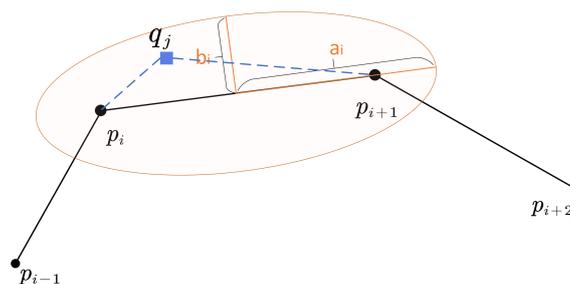


Figure 10. A anchor point and its matching ellipse.

In our method, every anchor point in a trajectory should have an impact on the overall similarity. The principles of designing the similarity model are:

- Every match between P and Q will have a positive effect on $S(P, Q)$.
- A continuous match will bring an additional increase in $S(P, Q)$.
- A mismatch has a non-positive effect on $S(P, Q)$, i.e., no effect or a negative effect.

Our proposed similarity measure based on the amended ellipse model is as follows:

$$s_{q_j} = \exp \left(- \min_{f \in \{p_i, p_{i+1}\}} \frac{d(q_j, f)}{a_i} \right) \cdot k(q_j) \quad (9)$$

$$k(q_j) = \begin{cases} \text{current length of continued match} & q_i \text{ is matched} \\ 0 & q_i \text{ is mismatched} \end{cases}$$

For every anchor point in both trajectories, we compute its contribution to the final similarity based on its position and its matched ellipse, like Figure 10 shows. The contribution of point q_j is denoted by s_{q_j} . Note that s_{q_j} contains a multiplication factor denoted by $k(q_j)$, and it is called the continuity coefficient, which is derived from the second principle. $k(q_j)$ is updated for each anchor point to record the current length of continued match. Its initial value is zero, and it will increase by one if the current point is matched to an ellipse from the other trajectory. It will be reset to zero when the current point is not matched to any ellipse. Afterwards, we can get the unnormalized similarity by summing up every anchor point's contribution:

$$S(P, Q) = \frac{1}{2} \left(\sum_{i=0}^{n-1} s_{q_i} + \sum_{j=0}^{m-1} s_{p_j} \right) \quad (10)$$

Since we multiply the single similarity measure by a continuity factor, we can normalize the overall similarity measure by dividing the maximum continuity:

$$NS(P, Q) = \frac{1}{2} \left(\sum_{i=0}^{n-1} s_{q_i} / \sum_{i=1}^{n-1} i + \sum_{j=0}^{m-1} s_{p_j} / \sum_{j=1}^{m-1} j \right) \quad (11)$$

$NS(P, Q)$ thus becomes our final UTSM. It can be easily proven from the formula definition that UTSM meets the requirements of being non-negative, symmetric, and reflexive.

4.3. Algorithms

We used two algorithms to compute the proposed similarity of uncertain trajectories based on our amended ellipse model. One is shape_estimator in Algorithm 1, which estimates the ellipse shape parameters according to the moving features computed by the position of each anchor point step-by-step. The other is UTSM_calculator Algorithm 2, which computes the UTSM value of a pair of trajectories by traversing the trajectory using two different loops. The overall computational complexity was $\mathcal{O}(pq)$, where p and q represent the length (i.e., the number of anchor points) of the two trajectories.

Algorithm 1: shape_estimator.

Input: $P \{(x_0, y_0, t_0), (x_1, y_1, t_1), \dots, (x_{n-1}, y_{n-1}, t_{n-1})\}$

Output: aList: a list of semi-major axis lengths of all $EL(p_i, p_{i+1}), i = 0, 1, \dots, (n-2)$

begin

 aList = [], $i = 1$

while $i < n-2$ **do**

$\vec{S}_i = (x_{i+1} - x_i, y_{i+1} - y_i)$ $\vec{S}_{former} = (x_{i+1} - x_{i-1}, y_{i+1} - y_{i-1})$

$\vec{S}_{latter} = (x_{i+2} - x_i, y_{i+2} - y_i)$ $\cos \theta_1 = \vec{S}_i \cdot \vec{S}_{former} / |\vec{S}_i| \cdot |\vec{S}_{former}|$

$\sin \theta_1 = \sqrt{1 - \cos^2 \theta_1}$ $\cos \theta_2 = \vec{S}_i \cdot \vec{S}_{latter} / |\vec{S}_i| \cdot |\vec{S}_{latter}|$ $\sin \theta_2 = \sqrt{1 - \cos^2 \theta_2}$

$\sin \alpha = 2 \times PoE \div |\vec{S}_i|$ $\cos \alpha = \sqrt{1 - \sin^2 \alpha}$ $v_1 = |\vec{S}_{former}| / (t_{i+1} - t_{i-1})$

$v_2 = |\vec{S}_{latter}| / (t_{i+2} - t_i)$ $b = [v_1 (\sin \theta_1 \times \cos \alpha + \cos \theta_1 \times \sin \alpha) +$

$v_2 (\sin \theta_2 \times \cos \alpha + \cos \theta_2 \times \sin \alpha)] / 8$ $a = \sqrt{b^2 + |\vec{S}_i|^2}$ aList.append(a) $i = i + 1$

Return aList

Algorithm 2: UTSM_calculator.

Input: a pair of trajectories P and Q; PoE: Positioning Error
Output: NS(P, Q): UTSM value of P and Q

begin

```

aListP = shape_estimator(P, PoE)
aListQ = shape_estimator(Q, PoE)
SP = 0
Qremain = Q
continuity = 0
/* compute the defined similarity from trajectory P to Q */
for p In P do
    /* traverse all the segments in trajectory Q */
    for (qi, qi+1) In Qremain do
        /* continuity auto-increment by default */
        continuity += 1
        if p is located In EL(qi, qi+1) then
            /* p is spatially located in the ellipse of segment (qi, qi+1) */
            SP += exp(-min(d(p, qi), d(p, qi+1))/aListQ[i]) × continuity
            /* next loop starts from the segment after the matched one */
            Qremain = Q[i+1: ]
            /* when find the matched ellipse, jump out of the current loop */
            Break
        else
            /* continuity reset to 0 when the current point mismatches */
            continuity = 0
    /* repeat the same operation from Q to P */
SQ = 0
Premain = P
continuity = 0
for q In Q do
    for (pi, pi+1) In Premain do
        continuity += 1
        if q is located In EL(pi, pi+1) then
            SQ += exp(-min(d(q, pi), d(q, pi+1))/aListP[i]) × continuity
            Premain = P[i+1: ]
            Break
        else
            continuity = 0
n = length(P), m = length(Q)
UTSM =  $\frac{1}{2} \left( S_P / \sum_{i=1}^{n-1} i + S_Q / \sum_{j=1}^{m-1} j \right)$ 
Return UTSM

```

5. Experimental Evaluation

The goals of the following experiments are:

1. evaluate the effectiveness of UTSM in measuring the similarity between trajectories.
2. validate the robustness of UTSM to outliers, different sampling rates, and asynchronous sampling.

We used both real-world data and synthetic data to test particular properties.

5.1. Experiment Setup

Real-world data:

1. Marine Cadastre: This maritime AIS dataset records approximately 30 attributes for 150,000 ships around the U.S. territory with a frequency of one GPS reading per two to 10 seconds. Each trajectory contained various numbers of anchor points ranging from 10 to 3000.
2. T-Drive dataset: This is comprised of bare-bone trajectories extracted from real track data of New York taxi trips. The sample data we used contained more than 20,000 trajectories. Each trajectory contained 100 to 500 anchor points.

These two datasets can be downloaded by the links in the Supplementary Materials at the end of the article.

Synthetic data:

1. Simulated trajectories with outliers: We added some outlier points to the original real-world trajectories
2. Simulated trajectories with different sampling rates: We carried out resampling operations with different intervals based on selected real trajectories.
3. Artificial asynchronous trajectories: We manually sampled points asynchronously from the sinusoidal curve at irreducible frequencies.

Environment:

Experiments were carried out using the Python 3.6 program on an Intel® Core™ i7-4710MQ CPU @ 2.50 GHz with 16 GB RAM and the Ubuntu 18.04 LTS operating system.

Baseline methods:

1. Distance-based measure: Hausdorff distance, Frechet distance, DTW [33], LCSS [34], ERP [35], EDR [36], and area-based distance [32].
2. Similarity measure considering uncertainty: simple-ellipse based similarity [4].

5.2. Evaluation and Discussion

5.2.1. Computational Efficiency

To evaluate the efficiency of the UTSM model, both Marine Cadastre and the T-Drive dataset were used as the input for our algorithms. UTSM similarity between a randomly selected trajectory with all the others was computed in the same environment. As the length of trajectories (i.e., the number of anchor points in trajectories) varied, the time consumption of computing UTSM spread between the scale of milliseconds and seconds. The overall result is shown in Figure 11.

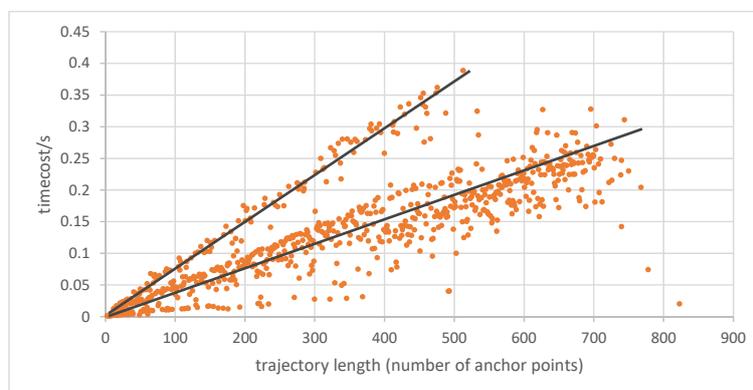


Figure 11. Uncertain Trajectory Similarity Measure (UTSM) computation time-cost of trajectories with different length.

It is obvious that the time-cost had a positive linear relationship with the length of the trajectories. To be more specific, there were two distinct linear trend lines of the scatter points. The trend line with a lower slope corresponded to the Marine Cadastre dataset, which contained a relatively smooth geometry of vessel trajectories. The other trend line with a higher slope corresponded to the T-Drive dataset, which contained automobiles' trajectories with more turns. It also indicated that the time-cost of computing UTSM was related to the geometry morphology resulting from the moving features of trajectories.

5.2.2. Effectiveness

We added different offsets to each anchor point of a real-world trajectory from the Marine Cadastre AISdataset and then computed the similarity between the original and the offset ones using different similarity measures. The results are shown in Table 2 and Figure 12. Note that to simulate the uniform similarity value, we conducted a conversion of the distance based value to transform it into similarity. The transformation formula is $S = 1 - e^{-\frac{1}{D}}$, where S represents the similarity and D represents the distance value. This was so that the distance based similarity could be transformed to the interval of [0, 1] no matter how large the distance value was.

Table 2. Similarity value of different measures for trajectories with different offsets.

Offset	5	10	20	50	100	200
Hausdorff	0.0855	0.0437	0.0221	0.0089	0.0044	0.0022
Frechet	0.0855	0.0437	0.0221	0.0089	0.0044	0.0022
DTW	0.0003	0.0001	7.4E-5	3.3E-5	1.8E-5	9.4E-6
LCSS	0.9906	0.9625	0.9031	0.8187	0.7562	0.6875
ERP	0.0003	0.0001	7.0E-5	2.8E-5	1.4E-5	7.0E-6
EDR	0.9906	0.9625	0.9125	0.8656	0.7968	0.7281
Area	4.1E-6	2.0E-6	1.0E-6	4.1E-7	1.0E-7	1.0E-7
Simple	0.4556	0.3235	0.1270	0.0566	0.0286	0.0171
UTSM	0.2831	0.2091	0.0020	0.0010	0.0004	0.0002

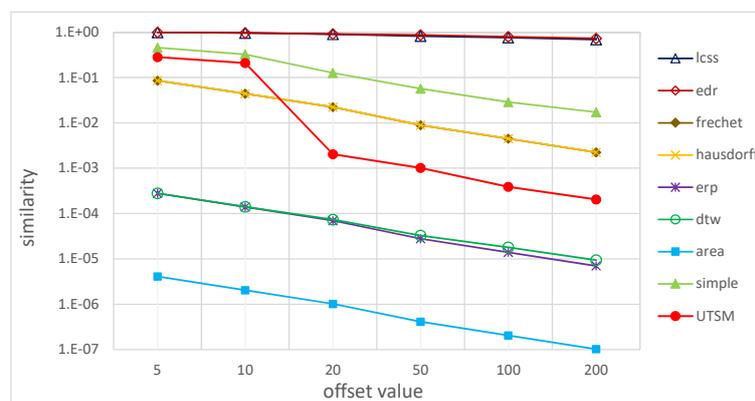


Figure 12. Similarities between trajectories with different offsets.

From the result, we can see that with the increase of manually added offset, all similarity measures dropped off. However, different measures started and dropped at different levels. When the offset was relatively small (i.e., five or 10), compared with the average distance between anchor points, some distance based measures, like area, ERP, DTW, Hausdorff, and Frechet, were relatively large, so that the corresponding similarity was too small to reflect the actual similarity of the sample trajectories. When the offset value was set bigger (i.e., 200), some measures such as LCSS, and EDR, remained big (>0.6), which indicated that they were not good similarity measures. Moreover, measures based on LCSS and EDR needed a reasonable threshold value in their algorithms to get a meaningful value, which restricts their application in many scenarios. In contrast, UTSM and the simple-ellipse

model started at a suitable value (i.e., 0.2 to 0.5) and showed a reasonable decreasing trend that could successfully reflect the actual similarity. This phenomenon also illustrated the effectiveness of considering uncertainty when measuring trajectory similarity. However, our proposed UTSM was more sensitive than the simple-ellipse model when the offset approached anchor points' interval of trajectories. The Marine Cadastre AIS dataset recorded the discrete position of vessel trips at a frequency of two to 10 seconds when the vessel was in motion, which means the distance between each pair of neighbor points in a vessel trajectory was about five to 30 meters (regarding the average speed of regular fishing or cargo ships). On this account, when we add an offset beyond that distance, the similarity value should decrease to a low level to reflect the actual characteristic of the trajectories. This is a strong evidence that our UTSM model could successfully capture the uncertain distance of the trajectories.

5.2.3. Robustness to Outliers

In this experiment, we selected relatively long trajectories from the T-Drive dataset and added noise of different sizes to several anchor points in the trajectory, like Figure 13 shows. The noise points were randomly selected from the trajectory, and this represented one percent of all anchor points.

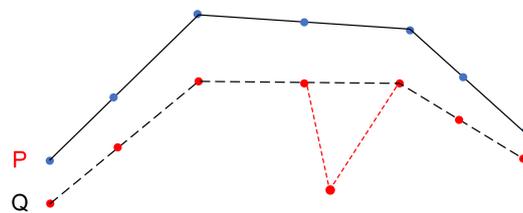


Figure 13. A trajectory with an outlier point.

To reflect the influence of the added outlier, we made a small offset to both the original and the noise trajectory and computed their similarity to the original one respectively. The two similarities are denoted as sim_{origin} and sim_{noised} . We also defined an Influence Ratio (IR) as $IR = (|sim_{noised} - sim_{origin}|) / (sim_{origin})$ to illustrate the impact degree of the added outlier. Figure 14 shows the IR result of different similarity measures.

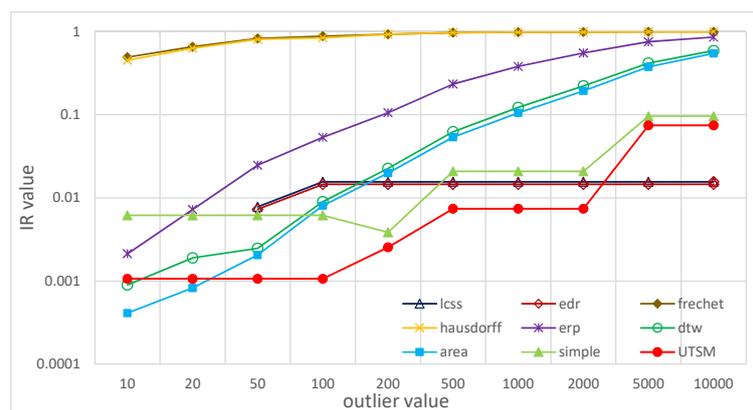


Figure 14. Influence Ratio (IR) under different outlier distances.

From the result, we can see that under different outlier sizes, the measurement models that consider uncertainty, i.e., the simple-ellipse model and our proposed UTSM, had smaller IR values compared with the other methods, especially when the added noise became large. UTSM performed even better than the simple-ellipse model, which indicated that UTSM had better robustness to outliers in the trajectory.

Noted that LCSS and EDR also had a small IR value under different noise sizes. This was because their input threshold value was set manually according to the added noise.

5.2.4. Tolerance to Different Sampling Rates

In this experiment, we again selected a relatively long trajectory from the T-Drive dataset and conducted a resampling operation on it with different intervals to simulate different sampling rates of trajectory data. Figure 15 is an example of the resampling operation.

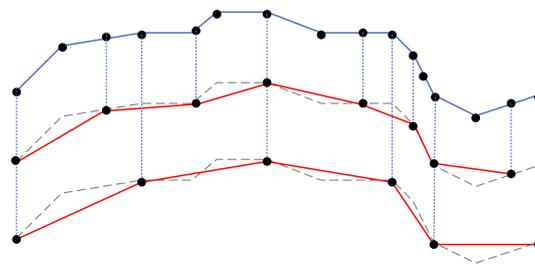


Figure 15. Resampling with intervals of two and three.

The same as the experiment in *V.B.3*, we also made a small offset to the original and the resampled trajectories and then computed the similarity with the original one. IR was also calculated to illustrate the influence of sampling frequencies on similarity models. The result is shown in Figure 16.

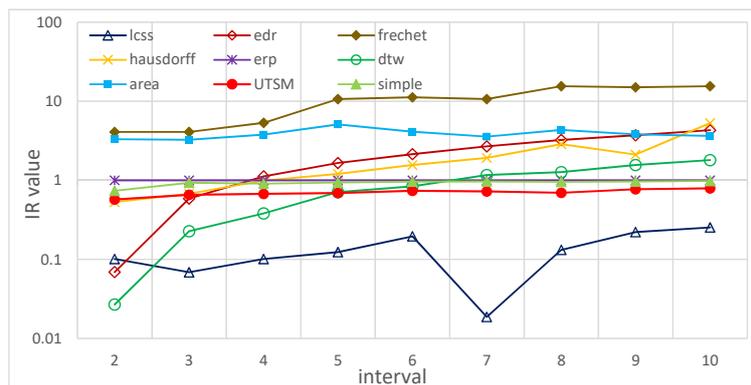


Figure 16. IR under different resample intervals.

From the result, we can see that our proposed UTSM model also showed good robustness to the sampling rate. The uncertain ellipse used in computing similarity could overcome the negative effect of missing data to a certain extent. UTSM had a relatively smaller IR value than the simple-ellipse model because UTSM took motion features into consideration. Its dynamically computed shape parameters could work better than simple ellipse when dealing with missing anchor points.

Noted that LCSS had the smallest IR value in most cases. This was because the LCSS model could neglect missing or mismatching points inherently when computing the similarity.

5.2.5. Tolerance to Asynchronous Sampling

Apart from the sampling rate, asynchronous sampling is also an important part of trajectory heterogeneity. In this experiment, we generated asynchronous trajectories by sampling on a continuous sinusoidal curve at different frequencies that were not divisible by each other. Figure 17 shows an example of two asynchronously sampled trajectories.

To evaluate the impact of asynchronous sampling, we took a fixed frequency (i.e., 100) as a benchmark and computed the similarity to trajectories sampled at other frequencies (i.e., 13, 23, 53, 83, 103, 123, 153, 203, 253, and 301), which were not divisible by the benchmark. The IR values of different measurement models are shown in Figure 18.

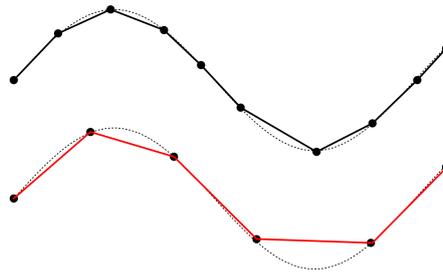


Figure 17. A pair of asynchronously sampled trajectories.

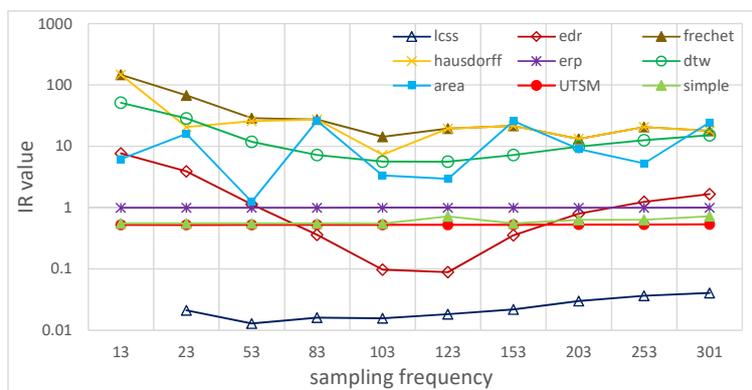


Figure 18. IR of asynchronous sampling trajectories.

From the result, we can observe that LCSS had the lowest IR value. This was also because LCSS’s algorithm can inherently neglect missing or unmatched points when computing similarity. This property contributed to better tolerance to different sampling strategies, but it had a negative effect on reflecting the actual similarity of trajectories, as experiment *V.B.2* demonstrated.

EDR had a low IR value when the sampling frequency was close to the benchmark frequency, but it performed worse if the frequency went farther. Therefore, basically, apart from LCSS, compared with other models, our proposed UTSM based on the amended ellipse had a lower IR value in most cases, which meant it showed better tolerance to asynchronous sampling.

5.2.6. Summary

To present the performance of different similarity measures in different scenarios more clearly, we summarize all the experiment results in Table 3. A measurement model is given a check mark if its performance ranks in the top three in the specific scenario.

Table 3 shows that our proposed UTSM was the only measure that ranked in the top three in all four evaluation indexes, which indicated that UTSM could be used as an effective and stable similarity measure in multi-source or heterogeneous trajectory datasets.

Table 3. Performance of different measures.

Measures	Effectiveness	Robustness		
		Outlier	Sampling Frequency	Asynchronous Sampling
Hausdorff	✓			
Frechet	✓			
DTW				✓
LCSS		✓	✓	✓
ERP				
EDR		✓		
Area				
Simple	✓		✓	✓
UTSM	✓	✓	✓	✓

6. Conclusions and Future Work

Our proposed amended ellipse model successfully addressed the uncertainty contained in trajectories. It took both interpolation error and positioning error into consideration. By computing the shape parameters of the ellipse dynamically based on motion features, each segment in a trajectory corresponded to an ellipse with different eccentricity, which described the uncertainty area of the segment adaptively. The proposed similarity measure UTSM based on the model could depict similarity in a normalized value, which could reflect the impact of uncertainty effectively. Compared with the work from [4], our amended model took the movement features into account when determining the shape parameters of the corresponding ellipse, rather than only using the maximum velocity. Besides, the influence of position error was considered in our model. When designing the UTSM similarity model, we emphasized the continuous match in the additional contribution of the final similarity value, which was also a significant improvement. Validated by experiments on both synthetic and real-world data, UTSM showed better robustness to noise and outliers than other classic measures. It was also more tolerant of different sampling rates and asynchronous sampling in trajectories. However, our proposed method had a relatively higher computational complexity because of the extra step when computing the parameters of the uncertainty ellipse.

For future work, the amended ellipse model can be further modified to adapt to time varying or conditional position error, and more research is needed to extend the uncertainty model to network constrained trajectories. For similarity measurement based on the uncertain model, we also plan to improve UTSM to make it compatible with non-equal interval sampling trajectory data and adopt the measurement to trajectory clustering applications. Furthermore, parallel algorithms need to be designed to improve the efficiency of similarity queries on large scale datasets. From the perspective of practical application, we plan to apply our proposed model in different trajectory related works, like trajectory similarity query, e.g., find certain flocks of birds that have a similar migration route, or map matching, e.g., match automobile or pedestrian trips to a certain part of the urban road network.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2220-9964/8/11/518/s1>. The real-world datasets that we used in the section “Experimental Evaluation” are available online. Marine Cadastre: <https://marinecadastre.gov/ais/>; T-Drive: <http://www.martinwerner.de/files/dataset-sample.tgz>.

Author Contributions: Ning Guo designed the model and implemented the algorithm; Ning Guo performed the validation experiments; Shashi Shekhar contributed to the preparation of the experimental data; Wei Xiong, Luo Chen and Ning Jing contributed to the building of the experimental environment; Ning Guo wrote the paper; Shashi Shekhar helped to improve the English language expression.

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. 41871284 and No. 41971362.

Acknowledgments: Thanks to the valuable comments and advice from Prof. Shashi Shekhar’s spatial computing group at the University of Minnesota.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Toohy, K.; Duckham, M. Trajectory similarity measures. *Sigspatial Spec.* **2015**, *7*, 43–50. [[CrossRef](#)]
2. Ranacher, P.; Tzavella, K. How to compare movement? A review of physical movement similarity measures in geographic information science and beyond. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 286–307. [[CrossRef](#)] [[PubMed](#)]
3. Yan, Z.; Parent, C.; Spaccapietra, S.; Chakraborty, D. A Hybrid Model and Computing Platform for Spatio-semantic Trajectories. In *The Semantic Web: Research and Applications*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 60–75.
4. Furtado, A.S.; Alvares, L.O.C.; Pelekis, N.; Theodoridis, Y.; Bogorny, V. Unveiling movement uncertainty for robust trajectory similarity analysis. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 140–168. [[CrossRef](#)]
5. GPS Product Team. *Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report*; GPS Product Team: Washington, DC, USA, 2014.
6. Reap, R.M. An Operational Three-Dimensional Trajectory Model. *J. Appl. Meteorol.* **1972**, *11*, 1193–1202. [[CrossRef](#)]
7. Tryfona, N.; Jensen, C.S. Conceptual Data Modeling for Spatiotemporal Applications. *Geoinformatica* **1999**, *3*, 245–268. [[CrossRef](#)]
8. Parent, C.; Spaccapietra, S.; Zimányi, E. Spatio-temporal conceptual models: Data structures + space + time. In Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, Kansas City, MO, USA, 2–6 November 1999; pp. 26–33.
9. Kuper, G.; Libkin, L.; Paredaens, J. *Constraint Databases*; Springer Science and Business Media: Berlin, Germany, 2013.
10. Spaccapietra, S.; Parent, C.; Damiani, M.L.; Macedo, J.A.D.; Porto, F.; Vangenot, C. A conceptual view on trajectories. *Data Knowl. Eng.* **2008**, *65*, 126–146. [[CrossRef](#)]
11. Hornsby, K.S.; Cole, S. Modeling moving geospatial objects from an event-based perspective. *Trans. GIS* **2007**, *11*, 555–573. [[CrossRef](#)]
12. Liu, H.; Schneider, M. Similarity measurement of moving object trajectories. In Proceedings of the Acm Sigspatial International Workshop on Geostreaming, Redondo Beach, CA, USA, 6 November 2012.
13. Wang, H.; Han, S.; Kai, Z.; Sadiq, S.; Zhou, X. An effectiveness study on trajectory similarity measures. In Proceedings of the Twenty-fourth Australasian Database Conference, Adelaide, Australia, 29 January–1 February 2013.
14. Ying, J.J.; Lu, E.H.; Lee, W.; Weng, T.; Tseng, V.S. Mining user similarity from semantic trajectories. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, San Jose, CA, USA, 2 November 2010; pp. 19–26.
15. Hornsby, K.; Egenhofer, M.J. Modeling Moving Objects over Multiple Granularities. *Ann. Math. Artif. Intell.* **2002**, *36*, 177–194. [[CrossRef](#)]
16. Hägerstrand, T. What about people in regional science? *Pap. Reg. Sci.* **1970**, *24*, 7–24. [[CrossRef](#)]
17. Miller, H.J. Modelling accessibility using space-time prism concepts within geographical information systems. *Int. J. Geogr. Inf. Syst.* **1991**, *5*, 287–301. [[CrossRef](#)]
18. Neutens, T.; Van de Weghe, N.; Witlox, F.; De Maeyer, P. A three-dimensional network-based space-time prism. *J. Geogr. Syst.* **2008**, *10*, 89–107. [[CrossRef](#)]
19. Miller, H.J. A measurement theory for time geography. *Geogr. Anal.* **2005**, *37*, 17–45. [[CrossRef](#)]
20. Downs, J.A.; Lamb, D.; Hyzer, G.; Loraamm, R.; Smith, Z.J.; O’Neal, B.M. Quantifying spatio-temporal interactions of animals using probabilistic space-time prisms. *Appl. Geogr.* **2014**, *55*, 1–8. [[CrossRef](#)]
21. Song, Y.; Miller, H.J.; Zhou, X.; Proffitt, D. Modeling Visit Probabilities within Network-Time Prisms Using Markov Techniques. *Geogr. Anal.* **2016**, *48*, 18–42. [[CrossRef](#)]
22. Kuijpers, B.; Miller, H.J.; Othman, W. Kinetic prisms: Incorporating acceleration limits into space-time prisms. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2164–2194. [[CrossRef](#)]
23. Pfoser, D.; Jensen, C.S. Capturing the uncertainty of moving-object representations. In *International Symposium on Spatial Databases*; Springer: Berlin, Germany, 1999; pp. 111–131.
24. Jeung, H.; Lu, H.; Sathe, S.; Yiu, M.L. Managing evolving uncertainty in trajectory databases. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 1692–1705. [[CrossRef](#)]

25. Frentzos, E.; Gratsias, K.; Theodoridis, Y. On the effect of location uncertainty in spatial querying. *IEEE Trans. Knowl. Data Eng.* **2008**, *21*, 366–383. [[CrossRef](#)]
26. Trajcevski, G.; Wolfson, O.; Hinrichs, K.; Chamberlain, S. Managing uncertainty in moving objects databases. *ACM Trans. Database Syst. (TODS)* **2004**, *29*, 463–507. [[CrossRef](#)]
27. Zhang, M.; Chen, S.; Jensen, C.S.; Ooi, B.C.; Zhang, Z. Effectively indexing uncertain moving objects for predictive queries. *Proc. VLDB Endow.* **2009**, *2*, 1198–1209. [[CrossRef](#)]
28. Song, Y.; Miller, H.J. Simulating visit probability distributions within planar space-time prisms. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 104–125. [[CrossRef](#)]
29. Miller, H.J. Time Geography and Space–Time Prism. In *International Encyclopedia of Geography: People, the Earth, Environment and Technology*; Wiley: Hoboken, NJ, USA, 2016; pp. 1–19.
30. Lee, J.; Miller, H.J. Analyzing collective accessibility using average space-time prisms. *Transp. Res. Part D: Transp. Environ.* **2019**, *69*, 250–264. [[CrossRef](#)]
31. Winter, S.; Yin, Z.C. The elements of probabilistic time geography. *GeoInformatica* **2011**, *15*, 417–434. [[CrossRef](#)]
32. Jekel, C.F.; Venter, G.; Venter, M.P.; Stander, N.; Haftka, R.T. Similarity measures for identifying material parameters from hysteresis loops using inverse analysis. *Int. J. Mater. Form.* **2019**, *12*, 355–378. [[CrossRef](#)]
33. Keogh, E.; Ratanamahatana, C.A. Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* **2005**, *7*, 358–386. [[CrossRef](#)]
34. Vlachos, M.; Kollios, G.; Gunopulos, D. Discovering Similar Multidimensional Trajectories. In Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, USA, 26 February–1 March 2002.
35. Chen, L.; Ng, R. On the marriage of lp-norms and edit distance. In Proceedings of the Thirtieth International Conference on Very large Data Bases-Volume 30, Toronto, ON, Canada, 31 August–3 September 2004; pp. 792–803.
36. Lei, C.; Özsu, M.T.; Oria, V. Robust and fast similarity search for moving object trajectories. In Proceedings of the Acm Sigmod International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005.
37. Morse, M.D.; Patel, J.M. An efficient and accurate method for evaluating time series similarity. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 12–14 June 2007; pp. 569–580.
38. Pelekis, N.; Kopanakis, I.; Marketos, G.; Ntoutsi, I.; Andrienko, G.; Theodoridis, Y. Similarity Search in Trajectory Databases. In Proceedings of the International Symposium on Temporal Representation and Reasoning, Alicante, Spain, 28–30 June 2007.
39. Dodge, S.; Laube, P.; Weibel, R. Movement similarity assessment using symbolic representation of trajectories. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 1563–1588. [[CrossRef](#)]
40. Naderivesal, S.; Kulik, L.; Bailey, J. An effective and versatile distance measure for spatiotemporal trajectories. *Data Min. Knowl. Discov.* **2019**, *33*, 577–606. [[CrossRef](#)]
41. Baccour, L.; Alimi, A.M.; John, R.I. Some notes on fuzzy similarity measures and application to classification of shapes recognition of arabic sentences and mosaic. *IAENG Int. J. Comput. Sci.* **2014**, *41*, 81–90.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).