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; guoning10@nudt.edu.cn, xiongwei@nudt.edu.cn, luochen@nudt.edu.cn, ningjing@nudt.edu.cn
2 Department of Computer Science, University of Minnesota, Minneapolis 55455, US; gning@umn.edu, 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 motion features of trajectory to compute the ellipse’s shape parameters. A specialized similarity measure method considering uncertainty called 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 is 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 based 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 it will remain accessible due to increasing privacy protection demands by 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 asynchronous sampling strategy will cause misjudgments. Consider the similarity comparison shown in Figure 1: the real trajectories of P and Q in Figure 1(a) are closer than those of P and R. But because of the low sampling frequency and heterogeneous distribution of the sampled points, the rebuilt trajectories in Figure 1(b) obtained by linear interpolation shows the opposite result: R is more similar to P than Q.

![Figure 1. A trajectory similarity distortion example](image)

To reduce this kind of misjudgment under 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] which 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 Uncertain Trajectory Similarity Measure (UTSM). Validated by experiments on both synthetic data and real-world data, UTSM shows better capability in comparing 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 to future work.

2. Preliminaries and Problem Definition

2.1. Basic Concepts

- **Spatiotemporal Trajectory**

  A spatiotemporal 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:

  \[
P = \{(s_0, t_0), (s_1, t_1), \ldots, (s_{n-1}, t_{n-1})\}
  \]

  where \(n\) is the number of sampling points of the entire trajectory and \(s_i\) is the position of the moving object at \(t_i\) time spot. 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 is 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 intensity 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 2 meters horizontally at a 95% confidence interval[5], 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-value 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. But
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:

\[
\begin{align*}
0 < S(P, Q) & \leq 1 \\
S(P, Q) & = S(Q, P) \\
S(P, P) & = 1
\end{align*}
\]  

(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                                        |
| $\Delta t$  | Time interval between two consecutive points                               |
| $d_{\text{euc}}(p,q)$ | 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 $Q$                                  |
| $\text{NS}(P, Q)$ | 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 is proposed in [6], which is operational to 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], MADS[8] and other featured models designed under particular application background 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 which 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 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], cylinder model [26], grid model [27], Bead model[15] (also called multi-granularities model) or ST-prism [17,19] model 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. But both of them ignore the positioning error of anchor points [4].

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]. But 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 Fretchet distance, DTW[33], LCSS[34], ERF[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 eliminates the maximum speed assumption in the original model and replaces it with an Approximate Upper Bound Distance. It then defines and calculates measures of alikeness, sharedness, and continuity. Finally, a comprehensive similarity expression is obtained. However, they simply expand 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 have 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 is based on a circular bounding area with a radius proportional to the estimated maximum speed. While this circular bounding area can tackle the
problem of asynchronous sampling to a certain degree, the maximum speed is 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 which have been applied in shape classification [41].

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

![Figure 3. Trajectory similarity measure methods](image)

To overcome this limitation of previous work, we propose a novel amended ellipse model which 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. But 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(pi, pi+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 reflected adjacent segments.

The interpolation error originates from the linear interpolation, the most commonly used interpolation method, 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

Figure 5. Motion features of a segment

Take the 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.:

\[
\begin{align*}
\theta_1^i &= | \angle (p_{i-1}, p_{i+1}) - \angle (p_i, p_{i+1}) | \\
\theta_2^i &= | \angle (p_i, p_{i+2}) - \angle (p_i, p_{i+1}) | \\
\end{align*}
\]  

(2)

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

It’s 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 \(v_{p_i} \sin \theta_k^i\) can result in a vertical distance of \(\frac{1}{2} v_{p_i} \sin \theta_k^i \cdot \Delta t\) until the component decelerates to zero. The deceleration time is set to \(\Delta t\) because the moving object has to return to anchor point \(P_{i+1}\). We can get two uncertain vertical distances derived from starting velocity and ending velocity of the segment. So we can estimate the value of the semi-minor axis \(b\) by computing the average of the two uncertain vertical distances:

\[
b = \frac{1}{2} \left( \frac{v_{p_i} \sin \theta_1^i \cdot \Delta t}{2} + \frac{v_{p_{i+1}} \sin \theta_2^i \cdot \Delta t}{2} \right)
\]  

\[
= \frac{1}{8} \Delta t \left( v_{p_i} \sin \theta_1^i + v_{p_{i+1}} \sin \theta_2^i \right)
\]  

(3)

Figure 6. Uncertain vertical range estimation based on turning angle

The amended ellipse parameters are thus:
\[ c = \frac{1}{2}d_{euclidean}(p_i, p_{i+1}) \]
\[ b = \frac{1}{8} \Delta t (|v_{p_i} \sin \theta_{1i}| + |v_{p_{i+1}} \sin \theta_{2i}|) \]
\[ a = \sqrt{b^2 + c^2} \]  

(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, this ‘straight’ kind of curves tend to have less similarity with each other because of the smaller area of uncertainty.

- 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, this ‘winding’ kind of curves have more tendency of a greater similarity with each other as a result of the bigger overlapping area of adjacent uncertainty extents.

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.

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

\[ \sin \alpha = \frac{\varepsilon}{\frac{1}{2}d} = \frac{2\varepsilon}{d} \]  

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

\[
\sin \alpha = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} \alpha^{2k+1} = \alpha - \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!} - \cdots \approx \alpha \quad (6)
\]

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

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

\[
c = \frac{1}{2} d_{\text{euc}}(p_i, p_{i+1}) = \frac{1}{2} d \\
b = \frac{1}{8}\Delta t \left[ v_{p_i} \sin \left( \theta^1_i + \frac{2\varepsilon}{d} \right) + v_{p_{i+1}} \sin \left( \theta^2_i + \frac{2\varepsilon}{d} \right) \right] \quad (7)
\]

\[
a = \sqrt{b^2 + c^2}
\]

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

\[
b = \frac{1}{8}\Delta t \left[ v_{p_i} \sin \left( \theta^1_i + \arcsin \frac{2\varepsilon}{d} \right) + v_{p_{i+1}} \sin \left( \theta^2_i + \arcsin \frac{2\varepsilon}{d} \right) \right] \quad (8)
\]

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 a anchor point \( p_i \) in P locates 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](image)

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 negative effect.

Our proposed similarity measure based on the amended ellipse model is as follows:
\[ s_{q_j} = \exp\left(-\min_{f \in \{p_i,p_{i+1}\}} \frac{d(q_j,f)}{a_i}\right) \cdot k(q_j) \]

\[ k(q_j) = \begin{cases} 
\text{current length of continuematch} & \text{if } q_j \text{ is matched} \\
0 & \text{if } q_j \text{ is mismatched} 
\end{cases} \]

(9)

For every anchor points in both trajectory, 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 \( s_{q_j} \). Noted that the \( s_{q_j} \) contains a multiplication factor denoted by \( k(q_j) \), it is called continuity coefficient which is derived from the seconed principle. \( k(q_j) \) is updated for each anchor point to record current length of continue match. Its initial value is zero and it will increase by one if current point is matched to a ellipse from the other trajectory. And it will be reset to zero when current point is not matched to any ellipse. Afterwards, we can get the unnormalized similarity by sum up every anchor point’s contribution:

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

(10)

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

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

(11)

\( NS(P,Q) \) thus becomes our final measure of uncertain trajectory similarity (UTSM). It can be easily proved from the formula definition that UTSM meets the requirements of being non-negative, symmetric, and reflexive.

4.3. Algorithms

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

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

5.1. Experiment Setup

**Real-World Data:**

1. Marine Cadastre: a maritime AIS dataset records approximately 30 attributes for 150,000 ships around the US territory with a frequency of one GPS reading per 2 to 10 seconds. Each trajectory contains various numbers of anchor points ranging from 10 to 3000.
Algorithm 1: shape_estimator

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

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

begin
    aList = [], $i = 1$
    while $i < n-2$
        $\vec{S}_i = (x_{i+1} - x_i, y_{i+1} - y_i)\vec{S}_{former} = (x_{i+1} - x_{i-1}, y_{i+1} - y_{i-1})$
        $\vec{S}_{latter} = (x_{i+2} - x_i, y_{i+2} - y_i)\cos \theta_1 = \frac{\vec{S}_i \cdot \vec{S}_{former}}{|\vec{S}_i| \cdot |\vec{S}_{former}|}$
        $\sin \theta_1 = \sqrt{1 - \cos^2 \theta_1}$
        $\cos \theta_2 = \frac{\vec{S}_i \cdot \vec{S}_{latter}}{|\vec{S}_i| \cdot |\vec{S}_{latter}|}$
        $\sin \theta_2 = \sqrt{1 - \cos^2 \theta_2}$
        $\sin \alpha = 2 \times PoE \div |\vec{S}_i| \cos \alpha = \sqrt{(1 - \sin^2 \alpha)}$
        $v_1 = \frac{\vec{S}_{former} / (t_{i+1} - t_{i-1})}{(t_{i+2} - t_i)}$
        $b = [v_1 (\sin \theta_1 \times \cos \alpha + \cos \theta_1 \times \sin \alpha)] +$
        $v_2 (\sin \theta_2 \times \cos \alpha + \cos \theta_2 \times \sin \alpha)] / 8$
        $a = \sqrt{b^2 + |\vec{S}_i|^2}$
        aList.append(a) $i = i + 1$
    end
Return aList

2. T-Drive dataset: bare-bone trajectories extracted from real track data of New York taxi trips. The sample data we used contains more than 20,000 trajectories. Each trajectory contains 100 to 500 anchor points.

These two datasets can be downloaded by the links in 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 Sinusoid curve at irreducible frequencies.

Environment:

Experiments were carried out using the Python 3.6 program on an Intel® Core™ i7-4710MQ CPU @ 2.50GH with 16 GB RAM and 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 UTSM model, both MarineCadastre and T-Drive dataset are used to be the input of our algorithms. UTSM similarity between a randomly selected trajectory with all the others is computed in the same environment. As the length of trajectories (i.e. the number of anchor points in trajectories) varies, the time consumption of computing UTSM spread between the scale of milliseconds and seconds. The overall result is shown in Figure 11.
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
    Q_remain = 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 (q_i, q_{i+1}) In Q_remain do
            /* continuity auto-increment by defualt */
            continuity += 1
            if p Locates In EL(q_i, q_{i+1}) then
                /* p is spatially located in the ellipse of segment (q_i, q_{i+1}) */
                SP += exp(-min(d(p, q_i), d(p, q_{i+1}))/aListQ[i]) \times continuity
                /* next loop starts from the segment after the matched one */
                Q_remain = Q[i+1: ]
                /* when find the matched ellipse, jump out of current loop */
                Break
            else
                /* continuity reset to 0 when current point mismatches */
                continuity = 0
                Q_remain = Q[i+1: ]
                Break
        end for
    end for

    /* repeat the same operation from Q to P */
    SQ = 0
    P_remain = P
    continuity = 0
    for q In Q do
        for (p_i, p_{i+1}) In P_remain do
            continuity += 1
            if q Locates In EL(p_i, p_{i+1}) then
                /* q is spatially located in the ellipse of segment (p_i, p_{i+1}) */
                SQ += exp(-min(d(q, p_i), d(q, p_{i+1}))/aListP[i]) \times continuity
                P_remain = P[i+1: ]
                Break
            else
                continuity = 0
                P_remain = P[i+1: ]
                Break
        end for
    end for

    n = length(P), m = length(Q)
    UTSM = \frac{1}{2} \left( \frac{SP}{\sum_{i=1}^{n-1} i} + \frac{SQ}{\sum_{j=1}^{m-1} j} \right)
    Return UTSM
```

It’s obvious that the time cost has a positive linear relationship with the length of trajectories. To be more specific, there are two distinct linear trendlines of the scatter points. The trendline with a lower slope corresponds to MarineCadastre dataset, which contains relatively smooth geometry of vessel trajectories. And the other trendline with a higher slope corresponds to T-Drive dataset, which contains automobile’s trajectories with more turns. It also indicates that the time cost of computing UTSM is related to the geometry morphology resulting from moving features of trajectories.

5.2.2. Effectiveness

We added different offsets to each anchor point of a real-world trajectory from MarineCadastre AIS dataset 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 uniformed similarity value, we conducted a conversion on the distance-based value to transform it into similarity. The transformation formula is $S = 1 - e^{-\frac{D}{D}}$, where $S$ represents the similarity and $D$ represents the distance value. So that the distance-based similarity can be transformed to the interval of $[0, 1]$ no matter how large the distance value is.

| 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 |

From the result, we can see that with the increase of manually added offset, all similarity measures drop off. But different measures start and drop at different levels. When the offset is relatively small (i.e., 5 or 10), comparing with the average distance between anchor points, some distance-based measures, like area, ERP, DTW, Hausdorff, Frechet, are relatively large, so that the corresponding similarity is too small to reflect the actual similarity of the sample trajectories. When the offset value is set bigger (i.e., 200), some measures such as LCSS, EDR remain big (> 0.6), which indicates that they are not a good similarity measure. Moreover, measures based on LCSS and EDR need 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 starts at a suitable value (i.e., 0.2 to 0.5) and show a reasonable decreasing trend that can successfully reflect the actual similarity. This
phenomenon also illustrates the effectiveness of considering uncertainty when measuring trajectory similarity. However, our proposed UTSM is more sensitive than the simple-ellipse model when offset approaching anchor points’ interval of trajectories. The Marine Cadastre AIS dataset records the discrete position of vessel trips at a frequency of 2 to 10 seconds when the vessel is in motion, which means the distance between each pair of neighbor points in a vessel trajectory is about 5 to 30 meters (regarding the average speed of regular fisher or cargo ship). 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. It is a strong evidence that our UTSM model can successfully capture the uncertain distance of the trajectories.

5.2.3. Robustness to Outliers

In this experiment, we select relatively long trajectories from the T-Drive dataset and add noise of different sizes to several anchor points in the trajectory, like Figure 13 shows. The noised points are randomly selected from the trajectory and the number is 1 per cent of all anchor points.

To reflect the influence of the added outlier, we make a small offset to both the original and the noised trajectory and computed their similarity to the original one respectively. The two similarities are denoted as $\text{sim}_\text{origin}$ and $\text{sim}_\text{noised}$. We also defined an Influence Ratio (IR) as 

$$IR = \left(\frac{|\text{sim}_\text{noised} - \text{sim}_\text{origin}|}{\text{sim}_\text{origin}}\right)$$

to illustrate the impact degree of the added outlier. Figure 14 shows the IR result of different similarity measures.

From the result, we can see that under different outlier sizes, measuring models that consider uncertainty, i.e., the simple-ellipse model and our proposed UTSM, have smaller IR values compared with other methods, especially when the added noise becomes large. And UTSM performs even better than the simple-ellipse model, which indicates that UTSM has better robustness to outliers in trajectory.

Noted that LCSS and EDR also have a small IR value under different noise sizes. This is because their input threshold value is set manually according to the added noise.
5.2.4. Tolerance to Different Sampling Rates

In this experiment, we again select a relatively long trajectory from the T-Drive dataset and conduct 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.

As the same with 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 is also calculated to illustrate the influence of sampling frequencies to similarity models. The result is shown in Figure 16.

From the result, we can see our proposed UTSM model also show good robustness to sampling rate. Cause uncertain ellipse used in computing similarity can overcome the negative effect of missing data to a certain extent. UTSM has relatively smaller IR value than simple-ellipse model because UTSM takes motion features into consideration. Its dynamically computed shape parameters can work better than simple ellipse when dealing with missing anchor points.
Noted that LCSS have the smallest IR value in most cases. This is because LCSS model can neglect missing or mismatching points inherently when computing the similarity.

5.2.5. Tolerance to Asynchronous Sampling

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

![Figure 17. A pair of asynchronously sampled trajectories](image)

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

![Figure 18. IR of asynchronous sampling trajectories](image)

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

EDR has low IR value when the sampling frequency is close to the benchmark frequency, but it performs worse if the frequency goes farther. So basically, apart from LCSS, compared with other models, our proposed UTSM based on the amended ellipse has lower IR value in most cases, which means it shows better tolerance to asynchronous sampling.
5.2.6. Summary

To present the performance of different similarity measures on different scenario more clearly, we summarize all the experiment results in table 3. A measure model is given a checkmark if its performance ranks top 3 in the specific scenario.

| Measures    | effectiveness | robustness |
|-------------|---------------|------------|
|             |               | outlier    | sampling frequency | asynchronous sampling |
| Hausdorff   | √             |            |                  |                      |
| Frechet     | √             |            |                  |                      |
| DTW         |               |            |                  |                      |
| LCSS        | √             | √          |                  | √                    |
| ERP         |               |            |                  |                      |
| EDR         | √             |            |                  |                      |
| Area        |               |            |                  |                      |
| Simple      | √             | √          | √                | √                    |
| UTSM        | √             | √          | √                | √                    |

Table 3 shows that our proposed UTSM is the only measurement that ranks top 3 in all the four evaluation indexes, which indicates that UTSM can be used as an effective and stable similarity measure in multi-source or heterogeneous trajectory datasets.

6. Conclusion and Future Work

Our proposed amended ellipse model successfully addresses the uncertainty contained in trajectories. It takes 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 corresponds to an ellipse with different eccentricity, which describes the uncertainty area of the segment adaptively. The proposed similarity measure UTSM based on the model can depict similarity in a normalized value which can reflect the impact of uncertainty effectively. Compared with the work from [4], our amended model takes the movement features into account when determining the shape parameters of corresponding ellipse, rather than only using maximum velocity. Besides, the influence of position error is considered in our model. When designing the UTSM similarity model, we emphasize the continuous match in the additional contribution of final similarity value, which is also a significant improvement. Validated by experiments on both synthetic and real-world data, UTSM shows better robustness to noise and outliers than other classic measures. It is also more tolerant of different sampling rates and asynchronous sampling in trajectories. However, our proposed method has a relatively higher computational complexity because of extra step when computing the parameters of 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 flock of birds that have the similar migration route, or map matching, e.g., match automobile or pedestrian trips to a certain part of urban road network.
**Supplementary Materials:** The real-world datasets that we use in 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 experiment environment; Ning Guo wrote the paper and 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. Toohey, K.; Duckham, M. Trajectory similarity measures. *Sigspatial Special 2015*, 7, 43–50.
2. Ranacher, P.; Tzavella, K. How to compare movement? A review of physical movement similarity measures in geographic information science and beyond. *Cartography and Geographic Information Science 2014*, 41, 286–307. doi:10.1080/15230406.2014.890071.
3. Yan, Z.; Parent, C.; Spaccapietra, S.; Chakraborty, D. A Hybrid Model and Computing Platform for Spatio-semantic Trajectories. The Semantic Web: Research and Applications; Springer Berlin Heidelberg: Berlin, Heidelberg, 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. *International Journal of Geographical Information Science 2018*, 32, 140–168. doi:10.1080/13658816.2017.1372763.
5. Team, G.P. Global positioning system (gps) standard positioning service (sps) performance analysis report, 2014.
6. Reap, R.M. An Operational Three-Dimensional Trajectory Model. *Journal of Applied Meteorology 1972*, 11, 1193–1202.
7. Tryfona, N.; Jensen, C.S. Conceptual Data Modeling for Spatiotemporal Applications. *Geoinformatica 1999*, 3, 245–268.
8. Parent, C.; Spaccapietra, S.; Zimányi, E. Spatio-temporal conceptual models: data structures + space + time. ACM International Symposium on Advances in Geographic Information Systems, Missouri, USA, 2-6 November 1999, pp. 26–33.
9. Kuper, G.; Libkin, L.; Paredaens, J. *Constraint databases*; Springer Science and Business Media, 2013.
10. Spaccapietra, S.; Parent, C.; Damiani, M.L.; Macedo, J.A.D.; Porto, F.; Vangenot, C. A conceptual view on trajectories. *Data and Knowledge Engineering 2008*, 65, 126–146.
11. Hornsby, K.S.; Cole, S. Modeling moving geospatial objects from an event-based perspective. *Transactions in GIS 2007*, 11, 555–573.
12. Liu, H.; Schneider, M. Similarity measurement of moving object trajectories. Acm Sigspatial International Workshop on Geostreaming, 2012.
13. Wang, H.; Han, S.; Kai, Z.; Sadiq, S.; Zhou, X. An effectiveness study on trajectory similarity measures. Twenty-fourth Australasian Database Conference, 2013.
14. Ying, J.J.; Lu, E.H.; Lee, W.; Weng, T.; Tseng, V.S. Mining user similarity from semantic trajectories. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. ACM, 2010, pp. 19–26.
15. Hornsby, K.; Egenhofer, M.J. Modeling Moving Objects over Multiple Granularities. *Annals of Mathematics and Artificial Intelligence 2002*, 36, 177–194.
16. Hägerstraand, T. What about people in regional science? *Papers in regional science 1970*, 24, 7–24.
17. Miller, H.J. Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information System 1991*, 5, 287–301.
18. Neutens, T.; Van de Wegen, N.; Witlox, F.; De Maeyer, P. A three-dimensional network-based space–time prism. *Journal of Geographical Systems 2008*, 10, 89–107.
19. Miller, H.J. A measurement theory for time geography. *Geographical analysis* 2005, 37, 17–45. doi:10.1111/j.1538-4632.2005.00575.x.

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. *Applied Geography* 2014, 55, 1–8.

21. Song, Y.; Miller, H.J.; Zhou, X.; Proffit, D. Modeling Visit Probabilities within Network-Time Prisms Using Markov Techniques. *Geographical Analysis* 2016, 48, 18–42.

22. Kuijpers, B.; Miller, H.J.; Othman, W. Kinetic prisms: incorporating acceleration limits into space-time prisms. *International Journal of Geographical Information Science* 2017, 31, 2164–2194.

23. Pfoser, D.; Jensen, C.S. Capturing the uncertainty of moving-object representations. *International Symposium on Spatial Databases*. Springer, 1999, pp. 111–131.

24. Jeung, H.; Lu, H.; Sathe, S.; Yiu, M.L. Managing evolving uncertainty in trajectory databases. *IEEE Transactions on Knowledge and Data Engineering* 2013, 26, 1692–1705.

25. Frentzos, E.; Gratsias, K.; Theodoridis, Y. On the effect of location uncertainty in spatial querying. *IEEE Transactions on Knowledge and Data Engineering* 2008, 21, 366–383.

26. Trajcevski, G.; Wollson, O.; Hinrichs, K.; Chamberlain, S. Managing uncertainty in moving objects databases. *ACM Transactions on Database Systems (TODS)* 2004, 29, 463–507.

27. Zhang, M.; Chen, S.; Jensen, C.S.; Ooi, B.C.; Zhang, Z. Effectively indexing uncertain moving objects for predictive queries. *Proceedings of the VLDB Endowment* 2009, 2, 1198–1209.

28. Song, Y.; Miller, H.J. Simulating visit probability distributions within planar space-time prisms. *International Journal of Geographical Information Science* 2014, 28, 104–125.

29. Miller, H.J. Time Geography and Space–Time Prism. *International encyclopedia of geography: People, the earth, environment and technologies*, 2016, pp. 1–19.

30. Lee, J.; Miller, H.J. Analyzing collective accessibility using average space-time prisms. *Transportation Research Part D: Transport and Environment* 2019, 69, 250–264.

31. Winter, S.; Yin, Z.C. The elements of probabilistic time geography. *GeoInformatica* 2011, 15, 417–434.

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. *International Journal of Material Forming* 2019, 12, 355–378.

33. Keogh, E.; Ratanamahatana, C.A. Exact indexing of dynamic time warping. *Knowledge and information systems* 2005, 7, 358–386.

34. Vlachos, M.; Kolios, G.; Gunopulos, D. Discovering Similar Multidimensional Trajectories. *Proc. of the 18th Int. Conf. on Data Engineering*, 2002.

35. Lei, C.; Ozsu, M.T.; Oria, V. Robust and fast similarity search for moving object trajectories. *ACM SIGMOD International Conference on Management of Data*, 2005.

36. Morse, M.D.; Patel, J.M. An efficient and accurate method for evaluating time series similarity. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007, pp. 569–580.

37. Naderivesal, S.; Kulik, L.; Bailey, J. An effective and versatile distance measure for spatiotemporal trajectories. *Data Mining and Knowledge Discovery* 2019, 33, 577–606.

38. 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 International Journal of Computer Science* 2014, 41, 81–90.