Program Name: NGA Academic Research Program (NARP)

Proposal Category: NURI

BAA Number: HM0476-18-BAA-0001

Title: Identifying Aberration Patterns in Multi-attribute Trajectory Data with Gaps

PI: Shashi Shekhar, Professor, University of Minnesota, Dept. of Computer Sc. and Eng.

Abstract:

We propose to investigate spatial data science approaches for Identifying Aberration Patterns in Multi-attribute Trajectory Data with Gaps (IAP-MTD). Example multi-attribute trajectory data (MTD) includes maritime MTD recording ship attributes (e.g., draught, rate of turn) and vehicle MTD recording on-board diagnostic attributes (e.g., emission). An aberration pattern represents a significant deviation from expected values. Identifying such aberration pattern can help improve maritime security and prevent illicit activities (e.g., illegal fishing, illegal oil transfer to violate United Nations sanctions) where the involved objects may hide their movement by deliberately not reporting their locations. The challenges of this problem arise from the complexity of modeling gaps and large amount of data. Existing works on trajectory mining focus on bare-bone trajectory data and consider only location-time information. In addition, they interpolate the gaps and ignore the many possibilities between consecutive reported locations. In contrast, to overcome the limitations in the literature, we propose a three-phase framework. First, we propose a novel frustum-chain model which represents multi-attribute trajectory data with gaps as well as the position measurement error of reported locations. Second, we propose query methods to efficiently discover aberration patterns with known spatiotemporal signatures. Third, we propose data mining approaches to discover aberration patterns without known spatiotemporal signatures. Both theoretical and experimental methodologies including proofs, complexity analysis, and experiments with synthetic as well as real datasets (e.g., MarineCadastre) will be used to evaluate the computational efficiency of the proposed methods. Furthermore, case studies will be used to evaluate effectiveness of proposed methods.
1. OBJECTIVES

We propose to investigate spatial data science approaches for Identifying Aberration Patterns in Multi-attribute Trajectory Data with Gaps (IAP-MTD). Examples of trajectory data with multiple attributes include maritime data with recordings of ship location, heading, draught, etc. An aberration pattern represents a significant deviation from expected values. An example is a long gap in a trajectory in a geographic area where a vast majority of trajectories do not have long gaps. IAP-MTD and aberration pattern may help detect malicious activities such as illegal cargo transfers, and illegal fishing.

IAP-MTD is challenging since gaps may cause traditional trajectory mining approaches [3] to underperform or fail as they assume the availability and preciseness of trajectory. Moreover, the problem is computationally challenging given the large data volume. For instance, MarineCadastre [1] is an Automatic Identification System (AIS) that records more than 30 attributes (e.g., location, draught) for 150,000 ships every minute. Its size reaches hundreds of Gigabytes as detailed in section 7.

The literature on movement pattern analysis [2] and trajectory mining [3, 4] interpolates trajectory gaps without considering the full range of an object’s movement possibilities. Furthermore, they focus on bare-bone trajectories containing only location information at each timestamp. Figure 1(a) shows a bare-bone trajectory of the movements of Napoleon’s 1812 Russian campaign army represented by a sequence of locations. By contrast, in Figure 1(b) the drawing by Charles Minard’s [5] shows a more informative multi-attribute trajectory for the same movements with attributes such as number of soldiers by thickness and temperature by notations along the trajectory. To overcome limitations of related work, we propose novel representation, querying, and mining approaches for MTD with gaps.

2. ANTICIPATED RESULTS

Anticipated results of this project include new mathematical models, computer algorithms, data analysis models, and analytical tools.

- Mathematical models: We anticipate developing a novel frustum-chain model to represent multi-attribute trajectories with gaps and position measurement error to uncover all possible movement directions between Reported locations.
- Computer algorithms: We expect to design new scalable algorithms (e.g., time-slicing and time prioritization) for data querying (e.g., spatial join) and data mining under the proposed frustum-chain model. These algorithms will need to scale to large datasets.
with billions of reported locations.

- Data analysis methods: We will develop new frustum-chain-based trajectory similarity measures which will consider location and attribute similarities as well as computational scalability to maintain applicability to large-scale data.
- Analytical tools: We will develop new software and extend the capabilities of existing software such as ArcGIS to facilitate a complete workflow from data input to actionable information incorporating the proposed frustum-chain model and algorithms.

3. APPLICABILITY

Marine safety and regulation enforcement are important for global security for concerns such as illegal oil transfer and transshipments. Figure 3 shows an example, where a Hong Kong ship possibly transferred oil to another ship from North Korea in the ocean, evading the United Nation sanctions [6] in October 2017. While performing these illicit activities, offenders tend to switch off their GPS transponders to hide their movement as described in a recent news article [7]: “When a ship goes dark in the Persian Gulf, it may be related to dodging sanctions”. Earlier in 2012, similar illegal transfer of oil to Iran possibly occurred in the Persian Gulf [8]. Transshipment is also a way to “dodge taxes” by avoiding higher tariffs while shipping goods between international subsidiaries [9].

Beyond global security concerns, many organizations and government agencies have put great effort in fighting against illegal fishing. For example, in 2014, the White House released the presidential memorandum “Establishing a Comprehensive Framework to Combat Illegal, Unreported, and Unregulated Fishing and Seafood Fraud” [10]. In Figure 2, a Panamanian commercial fishing vessel switched off location-reporting AIS device before entering the Galapagos Marine Reserve and then switched the device back on after 15 days possibly to hide illegal fishing [12]. “Data analysis found that of 200 Chinese vessels targeting squid off the coast of Peru, about 20% weren’t broadcasting a signal – indicating potential foul play.” [11].

4. APPROACH

We propose a three-stage research framework shown in Figure 4. The first stage focusses on representation of given input multi-attribute trajectory data (MTD) with gaps and position measurement error (PME) using a novel frustum-chain model that considers geometric properties (e.g., ship length and width) and laws in physics (details in Section
5.1). In the second stage, we will investigate approaches for querying this type of data to identify aberration patterns with known spatiotemporal (ST) signatures. Proposed querying approaches will define aberration patterns based on knowledge and theories of movement from variety of fields such as physics and behavioral science (details in Section 5.2). Stage three will be mining aberration patterns from MTD that does not have known ST signatures of aberration activities. The proposed approaches will discover novel patterns using statistical models which also leverage information provided by queries (details in Section 5.3).

![Diagram illustrating the three stages of the proposed approach]

**Fig 4. Three-stages of proposed approach**

5. SCIENCE

This section details the three proposed stages: (1) Representation of multi-attribute trajectory data (MTD) with gaps and position measurement errors (PME), (2) Querying such MTD with known spatiotemporal (ST) signatures of aberration activities, (3) Mining aberration patterns from such MTD without known ST signatures. Each stage includes a validation task to evaluate result quality and computational efficiency.

5.1 Representing MTD with gaps and Position Measurement Errors (PME).

In this task, we propose to develop a robust representation of MTD with gaps and PME. Traditionally, a multi-attribute trajectory can be represented as a sequence of ST points, each associated with a set of attributes. Figure 5(a) shows an example multi-attribute trajectory with the attribute of draught (represented by link thickness) in 1-D geospatial space for simplicity. When considering gaps between consecutive reported locations, the exact location of the object during the gap is unknown. However, it can be bounded by a candidate activity region (CAR). Figure 5(b) shows a space-time prism model [13] where each CAR is represented as a dashed parallelogram with a slope of 2, which is the assumed maximum speed of both objects. For simplicity, we only consider CARs of gaps (dashed line) as shown in Figure 5(c) and 5(d). In a 2-D geospatial space, each CAR is generalized from a parallelogram to the intersection between two cones that are vertexed at the start and end points of the gap, respectively as illustrated in Figure 6(a). The radii of the two cones at each timestamp depend on the maximum object speed and elapsed time [14]. In Figure 6(a), trajectory gap 1 (blue) starts at P\textsuperscript{1}\text{start} and ends at P\textsuperscript{1}\text{end}, each represented by two spatial coordinates and a temporal coordinate as summarized in the table in Figure 6(c). For example, P\textsuperscript{1}\text{start} is represented by spatial coordinates x\textsuperscript{1}\text{start}, y\textsuperscript{1}\text{start}, and temporal coordinate t\textsuperscript{1}\text{start}. Intuitively, starting at time point t\textsuperscript{1}\text{start}, the ship could move in the maximum speed S\textsuperscript{1}\text{max} towards any direction, which is modeled as a cone vertexed at P\textsuperscript{1}\text{start}. The radius of the cone at time t is \((t – t\textsuperscript{1}\text{start}) * S\textsuperscript{1}\text{max}. 
On the other side, since the object re-appears at $P_{\text{end}}$, the CAR has to be inside the cone vertexed at $P_{\text{end}}$ whose radius at time $t$ is $(t_{\text{end}} - t) \cdot S_{\text{max}}$. Thus, the CAR is the intersection between these two cones vertexed at $P_{\text{start}}$ and $P_{\text{end}}$ (blue).

However, the space-time prism [13] has limitations. It only considers the CAR between two consecutive reported locations and it ignores the PME of the reported locations themselves. Furthermore, it assumes that objects are points, which is unrealistic for large ships whose lengths range in hundreds of meters increasing uncertainty about ship’s geographic footprint inferred from a reported point location. We propose the following tasks to overcome limitations of the space-time prism model.

**Task R1 Model Position Measurement Error (PME):** We plan to develop a novel frustum-chain model in which PME is modeled as a circle centered at each reported location in a two-dimensional plane, with its radius representing the extent of error (e.g., standard deviation). For example, Figures 5(f) and 5(g) show the frustum-chain model in 1-D spatial space where the PME are represented as vertical lines in an one-dimensional geographical space. A CAR is now a hexagon with three parallel pairs of sides due to the fixed maximum speed assumption during a gap. When generalized to 2-D spatial space, in contrast of the intersection between two cones, the CAR will be formed by the intersection between two frustums between the start and end points of the gap. Figure
6(b) shows a CAR is intersected by two frustums, each starting at a circle centered at \( P_{1\text{start}} \) and \( P_{2\text{end}} \) with a radius of PME (i.e. \( r_{\text{PME}} \)). A trajectory with many gaps may be represented by a collection or chain of corresponding frustums. This model requires us to develop data structures to store the frustum-chain which is challenging due to the geometric complexity of frustums and the length of the trajectory.

**Task R2 Design a physics-aware model accounting for PME:** The model proposed in Task R1 is an overestimate of the CAR which ensures completeness but could lead to many false positives. Therefore, with the availability of data, we will also improve the precision of the proposed frustum-chain model by considering a richer motion model that accounts for acceleration [25] to assess the time needed for ships to either reverse or change movement direction. Figure 7 shows an illustration on 1-D spatial space. Compared to the original frustum-chain model on the left with a hexagonal candidate activity regions (CAR), the physics-aware model may have a more precise CAR bounded by two curves. As a simplified example, we assume the object has maximum speed \( S_{1\text{max}} \) at both start and end points \( P_{1\text{start}} \) and \( P_{1\text{end}} \) in opposite directions and can always achieve its maximum acceleration \( a_{1\text{max}} \). Then, the maximum distance travelled will be \( (t_{\text{end}}^{1} - t_{\text{start}}^{1}) \ast S_{1\text{max}} - S_{1\text{max}} \ast S_{1\text{max}} / a_{1\text{max}} \) which is shorter than that using the original frustum-chain model, \( (t_{\text{end}}^{1} - t_{\text{start}}^{1}) \ast S_{1\text{max}} \) since it takes time to gradually change the direction of movement and maximum speed cannot maintained all the time. Further, we will investigate physics-aware models that consider additional factors such as ship length and width, when available in the datasets such as the MarineCadaster [1].

**Task R3 Validation:** We will compare the proposed frustum-chain models in Tasks R1 and R2 with traditional space-time prism models for representation accuracy and computational costs (e.g., storage needs) through theoretical methods (e.g., theorems and proofs) and computer simulations.

### 5.2 Querying MTD with gaps and Position Measurement Error

Based on the frustum-chain model proposed in Section 5.1, we will next investigate approaches for querying MTD with trajectory gaps and PME. Specifically, we will develop queries for multiple maritime aberration patterns such as illegal fishing, illegal dumping, and illegal cargo transfer by exploiting known spatio-temporal (ST) signatures of such events. For example, the pattern depicted in Figure 5(h) is a signature of cargo transfer activity because the two CARs overlap (thick polygon) and the draught of ship 2 (blue) abruptly increases while that of ship 2 (red) abruptly decreases. It may be considered
aberration pattern if the cargo transfer occurs in open waters away from observers while not reporting ship locations.

Traditional spatial query processing methods process such ST join queries via a two-phase approach which treats each frustum as a 3-D ST object. First, a filter phase based on a 3-D R-tree [15] eliminates the candidate frustum pairs whose minimum bounding axis-parallel cuboids do not intersect and then a refine phase uses a plane sweep algorithm [16] to determine the exact intersection. However, these methods are inefficient in querying MTD with gaps due to the very high computational cost from a large number of ship trajectories, each modeled as a long frustum-chain. For example, if two frustums intersect very late in the chain, traditional plane sweep methods still need to traverse a large fraction of long frustum-chains before reaching the intersection. To overcome the limitations of traditional spatial query processing, we propose a three-phase efficient frustum-chain query approach as summarized in Figure 8. Their details are introduced in the following tasks.

Fig 8. Workflow of traditional approach and proposed approach.

**Task P1** Develop time-slicing and prioritization-based approach: We propose a novel prune-and-refine approach based on time-slicing and prioritization to efficiently query frustum-chains. The method starts with the same 3-D R-tree based filter as traditional methods as illustrated in Figure 8. For the surviving frustum pairs, we propose a new refine phase which slices each frustum along time where each time-slice is a snapshot modeled as the intersection between two circles. To determine whether two frustums intersect, we probe a novel priority-based traversal. For example, in Figure 5(h), we can start the traversal from the middle \( t = 4 \) of two synchronized gaps and immediately determine that the two gap frustums intersect. In contrast, a traditional plane sweep method will start at \( t = 2 \) then proceed to \( t = 3 \) and 4 until it detects the intersection and thus take more computations. We will investigate more complex scenarios such as unsynchronized gap frustums. Additionally, we will explore prioritization across gaps in a frustum-chain based on gap length and proximity to sensitive regions. For example, the portion of a long frustum-chain that passes a piracy-prone region or protected habitat may be prioritized.

**Task P2** Develop a more precise filter to query frustum-chains: Between the 3-D R-tree based filtering phase and the new refine phase, we will develop an intermediate filtering phase which bounds each frustum slice by a rectangle which is tighter than the corresponding slice of the 3-D R-tree, thereby achieving higher efficiency. Figure 9 shows an illustrative example where each 2D bounding rectangle (right) is smaller than the corresponding cross-section of 3D bounding cubic (left), leading to more efficient pruning.

**Task P3** Query multi-way aberrant rendezvous patterns with known ST signatures: The next task will be to extend queried objects from trajectory pairs to subsets of multiple trajectories which will help discover aberrant rendezvous among more than two objects
(e.g., multiple smaller oil tankers meeting up with a large oil tanker). To achieve high computational efficiency, we will investigate reducing redundant computations by modeling rendezvous patterns in an *apriori* [17] manner where a rendezvous involving $k$ objects can be defined based on its $k$ subset rendezvous each involving $k-1$ objects.

**Task P4 Validation:** We will evaluate the proposed time-slicing and prioritization-based approach as well as the more precise filter for querying frustum-chains theoretically and experimentally. Theoretical methods, e.g., proofs, will be used to characterize computational complexity, correctness and completeness. Experiments with synthetic and real datasets (e.g., *MarineCadastre* [1] detailed in section 7) will be used to evaluate the computational efficiency of the proposed models and algorithms under different combinations of parameters (e.g., dataset size, trajectory footprint).

### 5.3 Mining aberration patterns from MTD with gaps and PME

The goal of the third stage of our research is to discover new aberration patterns that have no known ST signature. State-of-the-art methods focus on bare-bone trajectory anomaly detection [18,19,20,21] which identifies deviations from usual movement corridors using either unsupervised clustering or supervised regression models to find frequently travelled corridors for different ship types. However, these approaches have major limitations when considering multi-attribute trajectories with gaps. First, they rely on similarity measurements (e.g., *Frechet* distance [22]) based on reported locations and fill trajectory gaps using linear interpolation, ignoring many movement possibilities covered by the proposed frustum-chain model. They also only consider the location trajectories of objects, and ignore other attributes (e.g., ship draught).

**Task M1 Develop similarity measures for frustum-chains:** We will first generalize the bare-bone trajectory anomaly detection framework to frustum-chains by defining novel similarity measures. For simplicity, frustum similarity may be measured by the fraction of the overlap time period. As shown in Figure 5(h), this is the ratio between the time range of the aberrant rendezvous event (i.e. approximately 3.5 to 4.9) and the time range of the entire frustum (i.e., 3 to 5). Richer frustum similarity measures may also account for the geographic area of possible rendezvous. Based on frustum pair similarity, we define the similarity between two frustum-chains $FC_1$ and $FC_2$ by aggregating the similarities of the frustum pairs formed by one frustum from $FC_1$ and one from $FC_2$. The main computational challenge comes from the large number of spatial joins of frustums to estimate overlap time and/or overlap area. To reduce computational cost, we propose to explore novel algorithms based on time-slicing and prioritization as discussed in Section 5.2.

In addition to measuring location-space similarity measure above, we will also model semantic similarity in frustum chains using information such as the length, frequency, and location of gaps. For example, an object may be aberrant if its trajectory has long gaps in a region where gaps rarely happen. Our semantic similarity will also
integrate information from attributes (e.g., ship draught). For instance, an illegal dumping while injecting back seawater could display an aberrant footprint because of the different densities of dumped material and sea water.

**Task M2 Develop aberration pattern detection approaches:** Based on the similarity measures proposed in Task M1, we will investigate aberration pattern detection approaches using an anomaly detection framework as follows: First, we will integrate location-time and semantic similarities into one unified similarity measure by assigning end-user provided weights to each type of similarities. Then, we will develop a clustering algorithm for finding normal frustum-chains and discover aberration patterns as those far from normal clusters. The second direction is to use Pareto frontier that processes the two types of similarity along two orthogonal axes and identifies non-dominated ships from a partial order among ships. Figure 10 shows an example Pareto chart where each dot represents a combination of two similarities. Then, aberration patterns can be discovered as points on the Pareto frontier, i.e., points not dominated by any other point on both similarities.

**Task M3 Develop classification and prediction models:** Once domain experts have confirmed the aberration patterns discovered by our methods, we will use these patterns and available ground-truth to explore methods of learning classification and prediction models to predict aberration patterns such as illegal fishing or illegal cargo transfer. We will investigate both supervised and semi-supervised models to overcome the challenges from limited availability of ground-truth.

**Task M4 Validation:** We will evaluate the proposed similarity measures for frustum-chain, aberration pattern detection approaches, as well as classification and prediction models theoretically and experimentally. Theoretical methods, e.g., proofs, will be used to characterize correctness and completeness. Controlled experiments will be conducted on synthetic and real datasets (e.g., MarineCadastre [1] detailed in section 7) to evaluate performance of the proposed approaches by varying parameters (e.g., dataset size, trajectory footprint). We will also use case studies to evaluate effectiveness of the proposed methods using cases with known ground truth.

**6. MANAGEMENT**

The PI, Dr. Shekhar, will provide overall leadership for the project. He has experience in leading similar and larger size research teams. This grant will enhance the capacities of the University of Minnesota in the geospatial analytics areas. A direct outcome of this research project will be the training and development of geospatial analysis students. Two Ph.D. students will be supported by research assistantships. Project outcomes will also be measured in terms of success in developing (i) robust representation of multi-attribute trajectory with gaps and PME, (ii) efficient querying of multiple maritime aberration patterns by exploiting known ST signatures, (iii) the effective discovery of aberration patterns without known ST signatures.
Schedule and Milestones: Table 1 shows the schedule of tasks. The first 6-month milestone is to represent 100 multi-attribute ship trajectories with gaps and position measurement errors using our frustum-chain model and to speed up queries on such data by 3 times over traditional methods. The second milestone is to speed up the algorithms for querying MTD with gaps and PME by 10 times. The third milestone is to speed up the aberration pattern mining algorithms by 10 times. The fourth milestone is to test the data mining approach with two case studies with available ground truth.

**Table 1: Project Task Schedule**

<table>
<thead>
<tr>
<th>Quarters</th>
<th>Year 1</th>
<th>Year 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>Q2</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representing MTD with gaps and PME</td>
<td>R1</td>
<td>R2</td>
</tr>
<tr>
<td>Querying MTD with gaps and PME</td>
<td>P1</td>
<td>P2</td>
</tr>
<tr>
<td>Mining aberration patterns from MTD with gaps and PME</td>
<td></td>
<td>M1</td>
</tr>
<tr>
<td>Tasks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Report</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the third, fourth and fifth option years we will further investigate physics aware model (R2) for improving precision in frustum-based model and explore classification approaches for aberration patterns. Additional research tasks and work-packages will be determined in consultation with NGA NURI program managers and GEO INT Analysts.

**7. FACILITIES AND EQUIPMENT**

We have access to open-source data from the MarineCadastre[1], an Automatic Identification System (AIS) that records more than 30 attributes (e.g., Maritime Mobile Service Identity, Longitude, Latitude etc.) for 150,000 ships around the US every minute. Its total size is about 600 Gb during 2009 to 2017. Its contents are summarized in Figure 11, which shows three entities, namely, Vessels, Trips and Location Signals. A vessel can voyage multiple trips and each trip can broadcast multiple location signals. Since MarineCadastre[1] data is limited to waters around the USA, we plan to purchase additional ship trajectory data from private vendors such as FleetMon [22], and Harris [23] for case studies near Galapagos Islands, Persian Gulf etc.

![Figure 11. Entity Relation Diagram of MarineCadastre[1] dataset](image)
The PI has access to computing facilities and resources within the University of Minnesota’s Department of Computer Science and Engineering and Minnesota Supercomputing Institute [26]. These include 741 nodes of various configurations with a total of 17,784 compute cores provided by Intel Haswell E5-2680v3 processors including 2 NVidia Tesla K20X GPUs with total memory of 67 TB. In addition, we have access to U-spatial [27] for spatial data, equipment and expertise. We will also purchase additional lab machines with high pixel count and high-resolution displays for data visualization.

8. SUB-AWARDS AND COLLABORATION

There are no sub-awards, however we are planning to collaborate with Prof. Harvey Miller at Ohio State University, an expert on space-time prism model [13]. Prof. Miller has contributed to modelling accessibility of space time prism which directly incorporate accessibility into locational analysis and transportation planning.

9. CURRENT AND PENDING SUPPORT

The research proposed in this proposal has not been and will not be submitted to any other sources of funding during the evaluation period.

10. RESEARCH TEAM DESCRIPTION

An ideal team for this project needs expertise in spatial data mining. Our team is experienced in spatial data mining with numerous publications [24]. Table 2 lists the names and of all persons for whom financial support is proposed, the planned commitments (in units of percentage of full-time work year) to the proposed research, and the planned commitments to other work and professional activities.

Table 2: Summary of Research Team and Planned Commitments

<table>
<thead>
<tr>
<th>Professional Commitments (During 9/2020 - 8/2025)</th>
<th>Annual Commitment (% full-time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI: Identifying Aberration Patterns in Multi-attribute Trajectory Data with Gaps (This Proposal).</td>
<td>8.33% 50%</td>
</tr>
<tr>
<td>Sr. Personnel (SP): Increasing low-input turfgrass adoption through breeding, innovation, and public education. (USDA/NIFA ends 8/21)</td>
<td>8.33%</td>
</tr>
<tr>
<td>PI: Spatio-temporal Informatics for Transportation Science (NSF: 8/1/19 – 7/31/23)</td>
<td>8.33%</td>
</tr>
<tr>
<td>Sr. Personnel (SP): Clinical and Translational Science Award (CTSA), National Institute for Health (NIH, 3/30/2018 - 2/28/2023).</td>
<td>5%</td>
</tr>
<tr>
<td>Co-P.I.: Midwest Big Data Hub: Building Communities to harness the data revolution. (NSF 06/01/2019 – 05/31/2023)</td>
<td>1%</td>
</tr>
<tr>
<td>University teaching and professional services</td>
<td>Balance Balance</td>
</tr>
<tr>
<td>Total</td>
<td>100% 100%</td>
</tr>
</tbody>
</table>
References


