Arun Sharma

Research Statement

Spatial data has tremendous value and is a necessary component in many critical societal applications. My research explores novel spatial data science and GeoAI techniques to address pressing societal challenges and develop innovative approaches to identify adversarial behaviors, particularly those involving data distortion. Current machine learning approaches frequently underperform because spatial data inherently violate the assumption of independent and identically distributed (i.i.d.) samples. Furthermore, even when these models perform reasonably well, they often fail to detect adversarial behaviors due to the scarcity of labeled training data and the absence of reliable ground-truth anomalies. Hence, my thesis explores novel and scalable spatial data science methods for imputing missing data that leverage known physics-based knowledge.

Keywords: Spatial Data Science, GeoAI, Spatial and Spatiotemporal Data Mining, Anomaly Detection.

1. Background

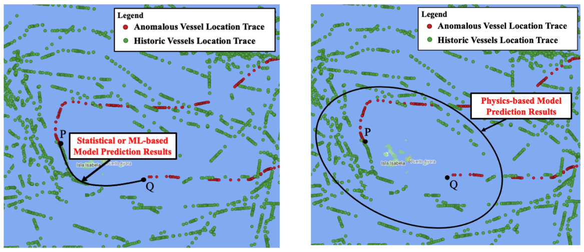
In recent years, spatial technologies such as Google Maps, Waze, Uber, Lyft, Grubhub, Lime, and autonomous driving have revolutionized our everyday lives. Location data from mobile devices generates hundreds of billions of dollars in revenue annually [1], with applications across sectors such as energy, healthcare, and retail. The world's economy also heavily relies on location and time data from billions of GPS receivers [2], which is essential to applications in banking, air travel, law enforcement, emergency services, and telecommunications. Meanwhile, new types of spatial data are emerging at unprecedented scale and volume. In transportation, a single vehicle can generate petabytes of onboard diagnostic data per hour [3]. Earth observation data is being collected at higher spatial resolution and frequency due to its value for transportation, infrastructure security, resource mapping, and agricultural monitoring. Spatial data research is funded by many National AI Institutes (AI-LEAF), and multiple Harnessing the Data Revolution (HDR) Institutes have been established. However, spatial data science and GeoAI face several unique challenges.

On the one hand, the knowledge derived from spatial data provides essential context for understanding and interpreting spatial patterns of events and objects across different areas. On the other hand, many scientific domains ignore or remain unaware of the special nature of spatial data. They largely adopt a spatial one-size-fits-all (OSFA) approach, where traditional data mining and machine learning methods are trained without accounting for the geographic properties that give rise to diverse geophysical and cultural phenomena. It is similar to choosing a general practitioner over a specialist when treating a complex medical condition, or to choosing multiple small language models for a specific task instead of a large foundation model. A key property of spatiotemporal data is that samples do not follow the independent and identically distributed (i.i.d.) assumption; instead, different geographic and temporal intervals exhibit distinct distributions. While many spatial data mining and statistical methods exist to address these issues, they tend to underperform when the data is manipulated, there are insufficient training samples, or the data is absent. Hence, my current area of interest is how to effectively handle missing or manipulated data, particularly in the context of intentional adversarial behavior, such as denial-based distortion (or denial-of-service) and spoofing.

2. Research Accomplishments

This research explores the broad area of denial-based distortion and GPS spoofing in spatial data science, aiming to understand the adversarial patterns generated through data distortion, which can indicate likely data manipulation, including missing-value imputations. Such adversarial pattern-detection methods require physical interpretation, and current machine learning methods often fail due to a lack of large training samples. Therefore, my thesis explores a physics-based approach to investigating adversarial behavior without prior training data. Understanding such data distortion caused by adversarial behavior is crucial to ensuring maritime safety and compliance with regulations [6, 7, 8, 9]. I addressed these challenges in two main lines of research: (1) a filter-and-refine strategy leveraging physics-based knowledge to identify anomalous trajectory gaps (i.e., denial-based distortion) and (2) identifying adversarial behavior using prior physical knowledge of motion characteristics with kinematic feasibility (i.e, deception). This work has resulted in multiple publications, including ACM TIST, ACM TSAS, AAAI, SIAM Data Mining (SDM), and ACM SIGSPATIAL.

2.1 Denial-based Distortion or Denial of Service: Physics-based Abnormal Trajectory Gap Patterns


Legend
Anomalous Vessel Location Trace
Historic Vessels Location Trace
Statistical or ML-based Model Prediction Results
P
Q
Isla Isabela
Puerto Ayora
Legend
Anomalous Vessel Location Trace
Historic Vessels Location Trace
Physics-based Model Prediction Results
P
Q
Isla Isabela
Puerto Ayora

(a) Trajectory Gap Prediction via Statistical or ML-based Models

(b) Trajectory Gap Prediction via Physics-based Model (Proposed Work)

Figure 2: Comparison of proposed physics-based with related work

(a) Filtering Phase: This phase is explicitly based on estimating geometric bounds for modeling adversarial patterns when there is no training data. Current data-driven methods fail to capture such adversarial behavior, making it reasonable to use a physics-based approach to model uncertainty. In this work, we first spatially quantify a trajectory gap based on the idea that an individual object's possible movements are constrained by certain limits of space and time. I represent these constrained movement possibilities as a space-time prism [11] by considering physics-based attributes (e.g., the object's max speed). This results in a more precise trajectory reconstruction technique than the classical assumption of the shortest or most frequently used path, as derived by current methods. The method captures specific out-of-sample predictions that are often missed by the shortest- or most-likely-path assumptions. For instance, Figure 2(a) shows statistical or machine learnng models trained on historical location data for the entire study area and outputs the most likely path an object would have taken while traveling from P to Q. By contrast, Figure 2(b) shows that we can leverage a simple physics-based model (e.g., a space-time prism) to constrain the object's movement in the form of a geo-ellipse, based on maximum speed (but unconstrained acceleration for simplicity) and without any prior knowledge of existing data [13].

Contributions: We first proposed a domain-specific explainable abnormal gap measure (AGM), defined as the probability that an object's location is not reported during a gap, given GPS coverage. The proposed AGM metric for quantifying anomalies is especially explainable in a maritime domain because ships have been legally required to report their location signals, and signal gaps that occur despite GPS coverage may indicate abnormal behavior. The method was also verified by a real-world case study of the Galapagos marine-protected habitat (Figure 2) [9], where the AGM scores based on a geo-ellipse captured denial-based abnormal behavior more accurately than linear interpolation. The explainability of these results enhances confidence in algorithmic reasoning, which is crucial for the responsible use of ML systems [12]. To further reduce redundant computations for scalability (e.g., enumerating trajectory gaps involved in intersections), we cached the derived polygonal areal coverage across multiple spatial joins and leveraged hierarchical indexing with space-time partitioning to simultaneously index and filter out non-intersecting gaps [13].

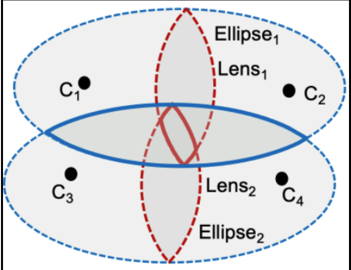


Fig 3: Lens Intersection at time t

(b) Refinement Phase: We further refined the derived geometric intersection of multiple physics-based geometric approximations by providing a tighter geometric filter at each temporal slice within each trajectory gap lifetime. At each time instant, we perform a circular intersection between the ellipse's foci, using the speed and time elapsed as the radius. The geometry defined by the two circular intersections is called a lens. For instance, in Figure 3, Ellipse$_1$ and Ellipse$_2$ denote two trajectory gaps derived via space-time prisms, and the intersection of the two ellipses shows a possible rendezvous area. However, at a given time instant *t*, the rendezvous area is the intersection of two lenses, Lens$_1$ and Lens$_2$.

(b) Contributions: We broadly addressed computational efficiency via time prioritization and a combination of static and dynamic filtering strategies. The time prioritizer narrows the time-slicing operation proposed in [14] by estimating the actual duration of the intersection between two space-time prisms. For a given spatial network with nodes (intersections) and edges (road segments), we also devised a Dual Convergence Trajectory Gap-Aware Rendezvous Detection (DC-TGARD) algorithm to scalably leverage the symmetry property of the geo-ellipse in the space-time prism by simultaneously filtering nodes in opposite directions (bi-directional pruning) until the lenses meet at the midpoint of the ellipse [16].

2.2 Deception or GPS Spoofing: Physics-informed Diffusion for Anomalous Trajectory Generation

GPS spoofing is defined as when an entity attempts to conceal its movements by intentionally emitting false signals (i.e., GPS spoofing) as shown in Figure 1(b) [17]. Detecting spoofing behavior remains challenging due to advances in AI-generated deep fakes (e.g., additive noise, fake trajectories). Moreover, the scarcity of ground-truth data on known GPS spoofing behavior hinders the development of accurate, physically plausible anomaly-detection methods. Finally, machine learning models provide statistical estimates and fail to incorporate physical constraints when generating trajectories. Current state-of-the-art methods overlook fine-scale spatiotemporal dependencies and prior physical knowledge, leading to lower accuracy. Hence, we proposed a physics-informed anomaly detection framework based on an encoder-decoder architecture that incorporates kinematic constraints to identify trajectories that violate physical laws.

4. Research Plan

4.1 Short-Term: Spatiotemporal agents (ST-Agents) built on foundation models (FMs) can adapt to diverse tasks via fine-tuning, few-shot, or zero-shot learning for dynamic spatial and cognitive reasoning. However, current data-driven methods require substantial computational resources, which increase non-renewable energy use [18], and are limited to transportation, not generalizing to other domains such as maritime environments. Integrating geographic context, spatial ontologies, and physical constraints into the state-of-the-art enables such ST-Agents to reason about spatial topology and semantics. Yet, data-driven STAgents struggle due to the scarcity of labeled data and computational demands. Thus, task-agnostic, physics- and knowledge-informed agents are needed for generalization. Pretraining embeddings with self-supervised objectives captures semantic and spatiotemporal dependencies, while modeling contextual cues (e.g., road networks, vessel traits, environmental conditions) boosts understanding and cuts costs.

To further mitigate computational overhead, ST-Agents can integrate surrogate models as lightweight approximations or microsimulators (e.g., SUMO, MATSim) to enable real-time pattern mining (e.g., anomaly detection). In addition, leveraging sophisticated physical models having historically rich theoretical foundations(e.g., Laws of Thermodynamics, Maxwell Equations), these surrogates emulate high-fidelity behaviors with reduced inference latency and energy demands, trained on subsampled data or pre-trained embeddings to preserve spatial reasoning accuracy while enabling real-time deployment with state-of-the-art orchestration frameworks (e.g., LangGraph). In urban/maritime domains, surrogates approximate constrained trajectory forecasts (e.g., hydrodynamics, traffic flows), supporting iterative optimization and uncertainty quantification at minimal cost. This curtails data-center carbon footprints and promotes edge computing scalability. Surrogate-assisted pretraining, via knowledge distillation from physics-informed FMs, bolsters cross-domain generalization toward energy-efficient spatial superintelligence.

4.2 Long-Term: One long-term task is to investigate challenges related to agent-based SpatialAI (e.g., ST-Agents), such as interpretability, robustness, bias, data quality, and adaptability. To enhance interpretability, future ST-Agents could incorporate techniques to trace and explain how specific domain knowledge influences autonomous decisions and predictions, fostering transparent agentic spatial cognition. To improve robustness, we must enable these agents to handle various uncertainties and environmental factors, such as noise, lighting variations, distribution shifts, and out-of-sample cases, while providing uncertainty quantification through adaptive reasoning mechanisms. Addressing biases will require ensuring that training data covers diverse scenarios, including clouds, polar regions, rare events, subpixel objects, and slow geological processes, to promote equitable spatial cognition across agents. Improving data quality will involve managing gaps, anomalies, and hotspots within dynamic spatiotemporal frameworks.

Finally, an open problem for Agent-based SpatialAI is how to achieve model generalizability across space while still allowing ST-Agents to capture spatial heterogeneity and evolve toward agentic spatial cognition. Given geospatial data with different spatial scales, we desire agents that can learn general spatial trends while still memorizing location-specific details through multi-scale reasoning. Will this generalizability introduce unavoidable intrinsic agent bias in downstream SpatialAI tasks? Will memorized localized information lead to an overly complicated prediction surface for a global prediction problem? Large-scale training data will likely exacerbate this problem, so careful consideration will be required. I am excited to make contributions in these areas over the next 5 to 10 years.

References:

[1] Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.
[2] “The World Economy Runs on GPS. It Needs a Backup Plan.” Bloomberg, Jul. 2018.
[3] “What’s driving the connected car?” McKinsey & Company, Sep. 2014
[4] Stoyanovich, Julia, Serge Abiteboul, Bill Howe, H. V. Jagadish, and Sebastian Schelter. "Responsible data management." Communications of the ACM 65, no. 6 (2022): 64-74.
[5] Stokes, Donald E. Pasteur's quadrant: Basic science and technological innovation. Brookings Institution Press, 2011.
[6] Goal 14: Conserve and sustainably use the oceans, seas, and marine resources, Sustainable Development Goals, United Nations.
[7] Launching the Grand Challenges for Ocean Conservation, World Wide Fund for Nature.
[8] Presidential Memorandum - Comprehensive Framework to Combat Illegal, Unreported, and Unregulated Fishing and Seafood Fraud, Office of Press Secretary, The White House, June 17, 2014.
[9] How Illegal Fishing Is Being Tracked From Space, Sarah Gibbens, National Geographic, December 3, 2018.
[10] Maritime Safety, International Maritime Organization, https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx
[11] Miller, Harvey J. "Time geography and space-time prism." International encyclopedia of geography: People, the earth, environment and technology 1 (2017).
[12] Sharma, Arun, Jayant Gupta, and Shashi Shekhar. "Abnormal Trajectory-Gap Detection: A Summary (Short Paper)." In 15th International Conference on Spatial Information Theory (COSIT 2022). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.
[13] Sharma, Arun, Subhankar Ghosh, and Shashi Shekhar. "Physics-based Abnormal Trajectory Gap Detection." ACM Transactions on Intelligent Systems and Technology.
[14] Sharma, Arun, Xun Tang, Jayant Gupta, Majid Farhadloo, and Shashi Shekhar. "Analyzing trajectory gaps for possible rendezvous: A summary of results." In 11th International Conference on Geographic Information Science (GIScience 2021)-Part I. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.
[15] Sharma, Arun, and Shashi Shekhar. "Analyzing Trajectory Gaps to Find Possible Rendezvous Region." ACM Transactions on Intelligent Systems and Technology (TIST) 13, no. 3 (2022): 1-23.
[16] Sharma, Arun, Jayant Gupta, and Subhankar Ghosh. "Towards a tighter bound on possible-rendezvous areas: preliminary results." In Proceedings of the 30th International Conference on Advances in Geographic Information Systems, pp. 1-11. 2022.
[17] Fake Signals and American Insurance: How a Dark Fleet Moves Russian Oil, Christiaan Triebert, Blacki Migliozzi, Alexander Cardia, Muyi Xiao and David Botti, NYTimes, May 30, 2023.
[18] How AI Is Fueling a Boom in Data Centers and Energy Demand, Andrew Chow, Time, June 12, 2024.