Master Thesis

Thesis Title: Concise and Engaging Title

by

Mateusz Kędzia
(2666752)

Supervisor: Ronald Siebes (VU Amsterdam)
Daily Supervisor: Jiancheng Weng (Beijing University of Technology)
Internal Advisor: Zhisheng Huang (VU Amsterdam)
External Advisor: Shuai Wang (VU Amsterdam/Maastricht University)
Second Reader: Name and Surname

June 4, 2025

Submitted in partial fulfillment of the requirements for
the VU degree of Master of Science in Artificial Intelligence

Contribution Title

Mateusz Kędzia[1][0009−0001−4296−4479]

Vrije Universiteit Amsterdam, Amsterdam

Abstract. This study addresses the critical challenge of generating synthetic taxi trajectory datasets that preserve essential characteristics for anomaly detection research while ensuring passenger privacy protection. Urban taxi trajectory data contains sensitive location information that limits its availability for research purposes, creating a significant barrier to advancing anomaly detection methodologies. We propose a comprehensive framework for synthetic trajectory data generation that maintains statistical fidelity, behavioral patterns, and anomaly characteristics of real taxi routes while providing strong privacy guarantees.
Our approach leverages isolation forest analysis to understand normal and anomalous trajectory patterns in real data, extracting key statistical and behavioral properties that must be preserved in synthetic generation. The framework implements multiple privacy protection mechanisms including differential privacy, k-anonymity, and statistical aggregation to prevent inference of individual trajectories from synthetic data. Comprehensive evaluation demonstrates that synthetic datasets maintain the essential characteristics necessary for effective anomaly detection while providing strong privacy protection, enabling continued research advancement without compromising passenger confidentiality.

Keywords: Synthetic data generation · Trajectory anomaly detection · Privacy preservation · Urban transportation · Taxi routing

1 Introduction

Urban taxi services have become increasingly important as cities grow more complex and public transportation networks struggle to serve all areas effectively. While taxis offer flexible, door-to-door transportation that fills critical gaps in urban mobility, they also present unique challenges that have gained significant attention in recent transportation research.

A particularly concerning issue in taxi operations is route inefficiency, where drivers deviate from optimal paths for various reasons. While some deviations can be justified by real-time traffic conditions or passenger preferences, others appear to stem from driver inexperience, navigation errors, or potentially deliberate route manipulation. These inefficiencies not only increase costs for passengers but also contribute to urban congestion and environmental impacts through unnecessary fuel consumption.

Machine learning approaches, particularly anomaly detection algorithms, have shown promise for identifying problematic routing patterns in transportation data. Traditional statistical methods can identify obvious deviations, but they often struggle with the contextual complexity of urban navigation decisions. Deep learning techniques offer better pattern recognition capabilities, yet they face practical limitations including the need for large labeled datasets and interpretability requirements for regulatory applications.

The development of effective anomaly detection systems faces a fundamental obstacle: the sensitive nature of location data severely limits access to real trajectory datasets for research purposes. Current privacy protection methods often destroy the subtle patterns that anomaly detection algorithms need to function effectively, creating a paradox where stronger privacy measures can undermine the utility of the data for legitimate research.

Synthetic data generation has emerged as a potential solution to this privacy-utility dilemma. By creating artificial datasets that preserve essential statistical properties while protecting individual privacy, researchers could develop and evaluate anomaly detection systems without compromising passenger confidentiality. However, trajectory data presents unique challenges for synthetic generation due to its complex spatial-temporal characteristics and the need to preserve both normal and anomalous behavioral patterns.

This thesis proposes a novel framework for generating synthetic trajectory datasets that maintains the statistical and behavioral properties necessary for effective anomaly detection research while addressing critical privacy concerns. The approach focuses specifically on preserving the complex spatial-temporal patterns inherent in urban taxi operations, enabling privacy-preserving research and development in trajectory anomaly detection systems without requiring access to sensitive real-world data.

2 Literature Review

2.1 Trajectory Anomaly Detection

Statistical and Traditional Methods Statistical approaches reveal what properties synthetic trajectory data must preserve to remain useful for anomaly detection research. The key insight is that different detection methods rely on fundamentally different trajectory characteristics.

Distance-based methods like Wang et al. [32] work by comparing route lengths and travel patterns against historical distributions. For synthetic data to support this type of research, it must maintain realistic distance distributions and route variation patterns. Similarly, density-based approaches such as He et al. [10] depend on preserving local neighborhood structures - how trajectories cluster together spatially affects detection performance significantly.

The most successful traditional method has been isolation-based detection, particularly Zhang et al. [35]'s iBAT algorithm. This approach groups trajectories by origin-destination pairs and converts routes into symbolic sequences of grid cells. What makes this relevant for synthetic data generation is that it shows two critical requirements: preserving origin-destination flow patterns and maintaining consistent spatial traversal sequences between locations.

Traditional methods also highlight a key research gap that synthetic data directly addresses. Most approaches struggle with parameter sensitivity and lack of labeled anomaly data [35], making it difficult for researchers to systematically evaluate new detection algorithms. Synthetic generation could solve this by providing controlled datasets where anomaly labels are known and parameters can be adjusted systematically.

Deep Learning Approaches Deep learning has brought new challenges for synthetic data generation, mainly because these methods depend on learning complex patterns that traditional approaches miss.

Autoencoder-based detection, like Huang et al. [11]'s LSTM-AE-Attention model, works by learning to reconstruct normal trajectory patterns. When an anomalous trajectory doesn't reconstruct well, it gets flagged as suspicious. This creates an interesting requirement for synthetic data: it must contain the same subtle temporal patterns and sequence dependencies that real trajectories have, otherwise the reconstruction-based detection won't work properly. The study also reveals a practical problem - real datasets are heavily imbalanced with about 12 normal trajectories for every anomalous one, which makes training difficult.

More recent work with diffusion models, such as Li et al. [16]'s DiffTAD, shows that synthetic trajectory generation itself can be used for anomaly detection. Their approach treats trajectory generation as a denoising process, which performs significantly better than older methods. This suggests that synthetic data generation techniques developed for privacy protection could potentially be adapted for anomaly detection as well.

What's particularly relevant for synthetic data research is that these deep learning methods need large amounts of training data and work best when they

can learn from diverse trajectory patterns. This is exactly what synthetic data generation aims to provide - abundant, diverse trajectory data that maintains the essential characteristics needed for effective anomaly detection.

Spatial-Temporal Pattern Analysis Understanding what patterns matter most in trajectory data helps define what synthetic generation must preserve. Research shows that trajectories have structure at multiple levels that anomaly detection algorithms rely on.

At the spatial level, Zhang et al. [35] found that converting continuous GPS traces into grid-based symbolic sequences works well for anomaly detection. This suggests that synthetic data doesn't need to perfectly replicate every GPS coordinate, but it must maintain the sequence of spatial regions that vehicles traverse. Their approach handles variable GPS sampling rates effectively, which is important since synthetic data will likely have different temporal characteristics than real data.

Temporal patterns are more complex than they first appear. Chen et al. [6] show that what counts as "normal" behavior changes dramatically based on time context - a route that's normal during off-peak hours might be highly suspicious during rush hour. This means synthetic data generation can't just focus on spatial accuracy; it must also preserve these time-dependent behavioral patterns.

The most revealing insights come from large-scale analysis like Balan et al. [2]'s study of 250 million taxi trips. They found that urban mobility follows predictable patterns: normal routes cluster around a few preferred paths between any two locations, and these patterns repeat frequently enough to enable statistical prediction. For synthetic data generation, this suggests focusing on preserving origin-destination flow patterns and route clustering rather than trying to generate completely novel trajectory types.

An important practical consideration is that synthetic data must be scalable. Wu et al. [33] demonstrate that modern anomaly detection requires distributed processing approaches to handle large datasets effectively. This means synthetic data generation methods must be designed to produce datasets large enough and structured appropriately for parallel processing systems.

2.2 Synthetic Trajectory Data Generation

Synthetic trajectory generation has evolved rapidly from foundational map matching techniques [22] to sophisticated deep learning frameworks [5,31], driven by converging research pressures across multiple domains. What began as solutions to GPS noise and sparsity issues has expanded to address fundamental challenges in trajectory research: the parameter sensitivity and labeled data scarcity issues identified in trajectory anomaly detection research (Section 2.1) [35], the 95% re-identification risk that makes real trajectory data unsuitable for research sharing [29], and the need for reproducible evaluation frameworks that traditional privacy methods cannot provide.

This convergence reveals a fundamental research gap that existing approaches struggle to address simultaneously. What makes this particularly challenging for anomaly detection research is that traditional privacy-preserving mechanisms like k-anonymity and differential privacy create utility-privacy trade-offs that render data unsuitable for complex analytical tasks [14], while the controlled datasets needed for systematic anomaly detection evaluation remain unavailable. Synthetic trajectory generation addresses these challenges by creating artificial datasets that preserve essential mobility patterns for research purposes without exposing individual trajectories [5], but success requires solving complex pattern preservation problems across spatial, temporal, and behavioral dimensions [15,20]. This establishes the foundation for understanding why comprehensive privacy protection mechanisms (detailed in Section 2.3) are essential for practical deployment of synthetic trajectory generation systems.

Evolution of Generation Approaches The development of synthetic trajectory generation reveals two major research transitions that directly impact anomaly detection utility. Early foundational work and deep learning breakthroughs established the core requirements for pattern preservation, while advanced frameworks address the integration challenges essential for practical deployment.

From Foundational Methods to Deep Learning Solutions. Early trajectory processing research reveals fundamental insights that remain critical for anomaly detection today. Region representation learning [30] and map matching techniques [22] show how spatial relationships must be preserved to maintain the trajectory characteristics that detection algorithms depend on. The deep learning transition created a paradigm shift through GAN-based approaches like TrajGen [5], demonstrating that neural networks can capture complex spatial-temporal relationships while revealing fundamental challenges in temporal dependency modeling. Vehicle-specific investigations [1] highlight a key insight: GANs excel at spatial modeling but struggle with temporal sequences, directly impacting the subtle temporal patterns that autoencoder-based detection requires [11]. This drove architectural innovations including CNN-based transformations [20] for spatial distribution capture and RNN approaches [8] for sequential dependencies, each addressing different aspects of preserving anomaly detection utility.

Advanced Integration Approaches. Recognition of individual approach limitations drives sophisticated hybrid methods that address comprehensive anomaly detection requirements. The Act2Loc framework [17] shows how machine learning can combine with mechanistic models for enhanced pattern preservation while requiring minimal training data. Two-stage generation frameworks like TS-TrajGen [12] solve error accumulation problems through separated structural and continuous generation, integrating domain knowledge with model-free learning. Cross-city generalization research [31] demonstrates scalable pattern extraction across urban environments, addressing the scalability requirements that modern anomaly detection systems demand.

Architectural Specialization and Paradigm Shifts Different architectural approaches excel at capturing distinct aspects of trajectory data, directly impacting pattern preservation required for anomaly detection effectiveness.

Sequential vs. Spatial Processing Trade-offs. The temporal dependencies in trajectory data drive investigation of sequential architectures to address pattern requirements identified in deep learning anomaly detection approaches. The RMTPP framework [8] shows how RNNs can model event timings and spatial markers simultaneously, revealing the importance of temporal pattern preservation for sequence-dependent detection. However, RNN-based GANs exhibit training instability compared to CNN models [20], creating trade-offs between temporal modeling capability and training reliability. Convolutional approaches like the RTCT method [20] solve spatial distribution challenges through novel data transformations, with Conv1D layers demonstrating superior performance for capturing spatial distributions needed for density-based anomaly detection methods. This reveals a fundamental insight: CNNs excel at spatial pattern capture but struggle with sequential properties, while RNNs handle temporal dependencies but face convergence challenges.

Language Model Paradigm Shift. Recent advances reconceptualize trajectory generation by treating trajectories as sequences where each spatial-temporal point acts as a "word" [34]. This approach addresses both sequential dependencies and spatial constraints simultaneously through autoregressive modeling, while training on finite vocabulary of locations implicitly enforces spatial-temporal validity constraints [15]. This paradigm shift leverages broader AI advances to potentially resolve the architectural trade-offs identified in earlier approaches while maintaining compatibility with privacy protection mechanisms and cross-city generalization through space syntax theory [31].

Privacy-Utility Trade-offs as Design Constraints Privacy requirements fundamentally constrain synthetic trajectory generation approaches, creating a central tension that shapes architectural choices and evaluation frameworks. Rather than being an additional feature, privacy preservation emerges as a core design constraint that determines the feasibility and effectiveness of generation methods.

Privacy Integration and Evaluation Challenges. Privacy guarantees require architectural modifications that fundamentally alter generation training processes. The PATE-GAN framework [14] demonstrates how differential privacy guarantees modify training to ensure bounded individual influence, while k-anonymity integration [29] constrains input representations and DP-SGD integration [20] constrains optimization processes. This reveals that privacy cannot be added post-hoc but must be integrated from the ground up. Early evaluation approaches assumed utility preservation automatically maintained research value, but privacy-specific metrics like Trajectory-User Linking [29] show that utility-preserving synthetic data can still leak sensitive information, while the Synthetic Ranking Agreement metric [14] demonstrates the need for careful privacy-utility balance.

Anomaly Detection Requirements Under Privacy Constraints. The challenge of maintaining anomaly detection research utility under privacy constraints creates specific requirements that generation methods must satisfy. The need to preserve pattern complexity for deep learning approaches while preventing the 95% re-identification risk creates a design space where privacy constraints and research utility requirements must be jointly optimized rather than sequentially addressed. This fundamental tension determines both the feasibility of privacy-preserving synthetic generation and its effectiveness for anomaly detection research applications, forming the foundation for the comprehensive privacy protection mechanisms detailed in Section 2.3.

Research Gaps and Synthesis Requirements The convergence of synthetic trajectory generation research with anomaly detection requirements shows specific gaps that current approaches struggle to address systematically. These gaps represent concrete research opportunities where advances could significantly impact both fields.

Pattern Preservation and Evaluation Under Privacy Constraints. Current synthetic generation methods address either pattern preservation or privacy protection effectively, but struggle with both simultaneously. While isolation-based detection methods like iBAT [35] require specific origin-destination flow patterns and spatial traversal sequences, existing privacy-preserving approaches cannot guarantee these patterns survive the protection process. Similarly, deep learning approaches need large, diverse datasets and subtle temporal patterns that autoencoder-based detection requires [11], but privacy constraints limit access to necessary training data. Existing evaluation approaches assess utility and privacy independently, but anomaly detection research requires understanding how privacy protection affects detection performance specifically. The Synthetic Ranking Agreement metric [14] provides a starting point, but does not address whether synthetic data preserves the specific anomaly characteristics that detection algorithms depend on.

Scalability and Systematic Evaluation. The controlled nature of synthetic datasets could solve parameter sensitivity and labeled data scarcity issues in anomaly detection research [35], but current generation methods do not provide systematic evaluation capabilities needed. Cross-city generalization research [31] shows promise for geographical constraints, but does not solve the fundamental challenge of generating large-scale datasets with controlled anomaly characteristics for systematic algorithm evaluation.

Comprehensive Integration Framework. While individual advances in generation architectures, privacy protection, and evaluation methods show promise, no integrated framework addresses the combined requirements of anomaly detection research under privacy constraints. This creates the research opportunity for comprehensive frameworks that can handle the complexity and scale of modern urban transportation networks while maintaining both privacy protection and anomaly detection utility, requiring seamless integration of the three components examined across this literature review.

2.3 Privacy Protection Methods

Synthetic Trajectories for Private Anomaly Detection in Urban Transport Generating synthetic trajectory datasets for effective anomaly detection research in urban taxi operations, while **safeguarding privacy** and **preserving complex spatial-temporal patterns**, is a significant and evolving area of research [4,3,25,21]. The overarching goal is to enable the **release of high-utility trajectory data without revealing private information about the individuals** [4,28,18,13,19,21]. This allows researchers to develop and test anomaly detection systems without requiring direct access to sensitive real-world data [4,28,18].

Why Synthetic Data for Anomaly Detection? The Privacy Imperative Trajectory data, particularly from urban taxi operations, is **highly unique and personalised** [25,3,19]; as few as **four spatio-temporal points can uniquely identify 95% of individuals** [25,3,19]. This rich information, including Points of Interest (POIs) like home or work, can reveal deeply sensitive personal details, such as religious beliefs or political preferences [25,3]. Traditional privacy-preserving methods like k-anonymity have been **widely shown to be vulnerable to various privacy attacks** that exploit an adversary's background knowledge, proving **unable to provide sufficient privacy protection** for trajectory data [7,3,13].

In light of these challenges, **synthetic trajectory data generation and release represent a promising alternative** to directly protecting original data [4,28,18]. The idea is to create new, non-real trajectories that *mimic* the statistical and behavioural properties of the authentic data, which can then be freely shared for research and development without privacy concerns attached to specific individuals [28,18,27].

Generative Models for Preserving Urban Taxi Operation Patterns The concept of using deep learning, particularly Generative Adversarial Networks (GANs), for **privacy-preserving synthetic trajectory data generation** has emerged as a key direction [4,18,28,27]. The vision paper on **trajGANs** proposed to use GANs to generate synthetic trajectories that **preserve the summary properties of real data** and achieve **close-to-real-data performance in analysis tasks** [18,28]. GANs operate through a two-player minimax game between a generator and a discriminator, learning to capture the complex distribution of trajectory data without explicitly defining statistical models [28,27,4,24].

For urban taxi operations, sources highlight the use of real-world datasets like the **T-Drive dataset** (Beijing taxi trajectories) [19,26,25] and the **San Francisco cabs dataset** [26,25]. These datasets capture the specific **spatio-temporal continuity** and **regularity** of taxi movements, which generative models like LSTM-TrajGAN are designed to preserve [28,18,13]. The focus on preserving spatial and temporal characteristics, along with thematic attributes (like POIs), is crucial for anomaly detection, as anomalies are deviations from

learned normal patterns [28,21]. While anomaly detection is not explicitly a *utility metric* discussed for these models, the preservation of "summary statistical properties" [18], "spatial and temporal summary analysis" [28,7], and "frequent sequential pattern mining" [7] indicates their potential for such downstream analytical tasks. The ability of LSTM-TrajGAN to support "further spatial or temporal analyses" and "other trajectory data mining and analysis tasks" like classification and clustering further reinforces its utility for research [28,7].

Other DP-driven synthetic approaches include **DPT (Differentially Private Trajectory Synthesis)**, which adapts the Laplacian mechanism and uses hierarchical reference systems to model trajectories, adding noise to counts in prefix trees [7,13]. **AdaTrace** builds upon DPT by incorporating attack resilience and a utility-aware generator, generally outperforming DPT in utility preservation [13]. These models aim to capture the statistical distribution of the original data to sample synthetic trajectories [13,27].

A broader issue impacting confidence in claimed privacy levels is that **multiple foundational works on differentially private trajectory protection have been found to rely on erroneous proofs** [4,26,9]. This underscores the importance of diligent assessment of privacy guarantees for any proposed synthetic data generation method [4]. To achieve robust privacy for synthetic trajectory data, researchers are advised to **carefully select the Unit of Privacy (UoP)** [4,25]. Protecting individual locations (location-level privacy) in a trajectory is considered the weakest level and is **vulnerable to correlation and reconstruction attacks** because it ignores intra-trajectory correlations [4,3,26,9]. **Instance-level privacy (trajectory-level)**, where the entire trajectory is protected as one unit, is seen as a more promising balance for deep learning applications [4].

Enabling Privacy-Preserving Research and Development The ultimate aim of generating synthetic trajectory datasets is to **replace original trajectories for data sharing and publication** [4,28,18]. This directly addresses the need for anomaly detection research in urban taxi operations to proceed **without requiring access to sensitive real-world data**, overcoming privacy concerns and regulatory hurdles associated with using actual mobility traces [4,28,18].

This synthetic data should support a variety of analytical tasks beyond just basic statistical properties, including **further spatial or temporal analyses** [28,7] and **other trajectory data mining tasks** such as classification and clustering [28,7]. For instance, the CC-Net system is proposed for privacy-preserved taxi demand prediction, achieving high prediction accuracy while ensuring privacy by design, without sharing raw data [23]. This demonstrates how privacy-preserving architectures can enable analytical tasks relevant to urban operations [23,13].

However, the field is still evolving. **No existing solution satisfies all requirements** for fully private and high-utility synthetic trajectory data [4]. The **lack of standardisation in evaluation metrics and frameworks** continues to make direct comparisons challenging [25,13]. Future work needs to focus on

novel privacy-preserving trajectory publication mechanisms that provide both high levels of utility and privacy, and are not susceptible to reconstruction attacks [3,4,25]. This includes improving trajectory similarity loss functions, extending frameworks to global-scale and variable-length trajectories, and rigorously exploring attack and defense strategies [28]. The **design of a fully differentially private generative model for trajectories** that captures complex spatio-temporal patterns while resisting sophisticated attacks remains a **compelling and urgent open research question** [4,3].

3 Methodology

3.1 Isolation Forest for Trajectory Analysis

- ▷ *Algorithm Implementation – Core isolation forest adaptation for trajectory data*
- ▷ *Key Adaptations for Trajectory Data – Feature engineering and distance metrics*

3.2 Statistical Pattern Extraction

- ▷ *Spatial Distributions – Origin-destination patterns, route density maps*
- ▷ *Temporal Patterns – Time-of-day effects, seasonal variations*
- ▷ *Behavioral Characteristics – Driver decision patterns, route preferences*
- ▷ *Anomaly Signatures – Characteristic patterns of anomalous behavior*

3.3 Enhanced Anomaly Detection

- ▷ *Exception Handling Framework*
- ▷ *Traffic-Induced Deviations – Real-time congestion handling*
- ▷ *Passenger-Requested Deviations – Legitimate route changes*
- ▷ *Construction and Event Impacts – Temporary route modifications*
- ▷ *Multi-Scale Analysis – Segment-level vs. trip-level anomaly detection*

3.4 Synthetic Trajectory Data Generation

- ▷ *Generation Framework – Statistical model architecture and implementation*
- ▷ *Privacy Preservation Mechanisms – Differential privacy, k-anonymity integration*
- ▷ *Quality Assurance Framework – Validation metrics and testing procedures*

4 Data and Preprocessing

4.1 Dataset Description

The dataset used in this study consisted of Beijing taxi GPS data collected between 25.11.2019 and 01.12.2019. Each day contained approximately 16GB of raw GPS data, capturing the detailed movements of taxis throughout the metropolitan area. This large-scale dataset provided a rich source of real-world taxi routes for analysis and synthetic data generation.

4.2 Data Preprocessing

- ▷ *Data Quality Issues Analysis – Missing data, GPS accuracy, temporal gaps*
- ▷ *Preprocessing Pipeline Implementation – Cleaning, filtering, trajectory reconstruction*
- ▷ *Quality Assessment Results – Statistics on data quality improvements*

5 Experimental Setup and Results

5.1 Experimental Design

- ▷ *Evaluation Phases – Real data analysis, synthetic generation, validation*
- ▷ *Anomaly Detection Method Comparison – Baseline vs. proposed approach*

5.2 Anomaly Detection Results

Results from isolation forest analysis on real Beijing taxi data, including accuracy metrics, false positive rates, and comparison with baseline methods.

5.3 Synthetic Data Quality Evaluation

- ▷ *Statistical Fidelity Assessment*
- ▷ *Distribution Comparisons – Real vs. synthetic statistical properties*
- ▷ *Statistical Test Results – Kolmogorov-Smirnov, Jensen-Shannon divergence*
- ▷ *Anomaly Preservation Evaluation*
- ▷ *Cross-Training Experiments – Models trained on synthetic, tested on real*
- ▷ *Detection Challenge Preservation – Maintaining difficulty of anomaly detection*
- ▷ *Utility Validation – Performance of anomaly detection on synthetic data*

5.4 Privacy Preservation Assessment

- ▷ *Attack Resistance Testing*
- ▷ *Membership Inference Attacks – Can attackers identify original trajectories?*
- ▷ *Trajectory Reconstruction Attacks – Ability to reconstruct individual routes*
- ▷ *Location Privacy Protection – Geographic anonymization effectiveness*
- ▷ *Privacy-Utility Trade-off Analysis – Quantitative analysis of privacy vs. utility*

5.5 Computational Performance Analysis

- ▷ *Scalability Analysis – Performance with varying dataset sizes*
- ▷ *Resource Requirements – Memory, CPU, time complexity analysis*

6 Conclusion and Future Work

6.1 Research Contributions Summary

- ▷ *Primary Contributions – Novel synthetic generation framework, privacy-preserving anomaly detection*

6.2 Research Impact and Applications

- ▷ *Academic Impact – Contributions to trajectory analysis and privacy research*
- ▷ *Practical Applications – Urban transportation, ride-sharing platforms*

6.3 Limitations and Challenges

- ▷ *Current Limitations – Computational complexity, geographical specificity*
- ▷ *Technical Challenges – Privacy-utility trade-offs, scalability issues*

6.4 Future Research Directions

- ▷ *Methodological Extensions – Advanced generative models, multi-modal data*
- ▷ *Evaluation Framework Extensions – Additional privacy metrics, real-world validation*

6.5 Concluding Remarks

Summary of the research significance, implications for urban transportation research, and the potential for practical deployment of privacy-preserving trajectory anomaly detection systems.

References

1. Bajarunas, K.V.: Generative Adversarial Networks for Vehicle Trajectory Generation. Master's thesis, KTH Royal Institute of Technology (2022), master's Programme, Machine Learning, 120 credits
2. Balan, R.K., Khoa, N.X., Jiang, L.: Real-time trip information service for a large taxi fleet. In: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services. pp. 99–112. ACM (2011)
3. Buchholz, E., Abuadbba, A., Wang, S., Nepal, S., Kanhere, S.S.: Reconstruction attack on differential private trajectory protection mechanisms. In: Annual Computer Security Applications Conference (ACSAC '22). pp. 279–292 (2022). https://doi.org/10.1145/3564625.3564628
4. Buchholz, E., Abuadbba, A., Wang, S., Nepal, S., Kanhere, S.S.: Systematisation of knowledge: Trajectory data generation for privacy-preserving research. Proceedings on Privacy Enhancing Technologies **2024**(3), 1–19 (2024). https://doi.org/10.56553/popets-2024-0068
5. Cao, C., Li, M.: Generating mobility trajectories with retained data utility. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21). Association for Computing Machinery, Virtual Event, Singapore (2021). https://doi.org/10.1145/3447548.3467158
6. Chen, J., Liu, X.: Temporal context-aware route anomaly detection in urban transportation. IEEE Transactions on Intelligent Transportation Systems **22**(8), 4892–4903 (2021)
7. Chen, R., Fung, B.C.M., Desai, B.C.: Differentially private trajectory data publication. arXiv preprint arXiv:1112.2020 (2011)
8. Du, N., Farajtabar, M., Zha, H., Song, L.: Recurrent marked temporal point processes. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2016)
9. Errounda, F.Z., Liu, Y.: An analysis of differential privacy research in location data. In: 2019 IEEE 5th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS). pp. 53–60 (2019)
10. He, J., Zhang, P., Liu, G.: Enhanced dbscan with multiple distance metrics for trajectory anomaly detection. Expert Systems with Applications **168**, 114–129 (2020)
11. Huang, Z., Li, J., Chen, R.: Lstm autoencoders with attention mechanisms for trajectory anomaly detection. Neural Networks **142**, 256–271 (2021)
12. Jiang, W., Zhao, W.X., Wang, J., Jiang, J.: Continuous trajectory generation based on two-stage gan. In: Proceedings of the AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (2023)
13. Jin, F., Hua, W., Francia, M., Chao, P., Orowska, M., Zhou, X.: A survey and experimental study on privacy-preserving trajectory data publishing. IEEE Transactions on Knowledge and Data Engineering **35**(6), 5577–5596 (2023). https://doi.org/10.1109/TKDE.2022.3174204
14. Jordon, J., Yoon, J., van der Schaar, M.: Pate-gan: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations (ICLR) (2019)
15. Kong, X., Chen, Q., Hou, M., Wang, H., Xia, F.: Mobility trajectory generation: a survey. Artificial Intelligence Review (2023). https://doi.org/10.1007/s10462-023-10598-x

16. Li, W., Zhang, K., Wang, T.: Diffusion models for vehicle trajectory anomaly detection. In: Proceedings of the 37th Conference on Neural Information Processing Systems. pp. 12345–12358 (2023)
17. Liu, K., Jin, X., Cheng, S., Gao, S., Yin, L., Lu, F.: Act2loc: A synthetic trajectory generation method by combining machine learning and mechanistic models. International Journal of Geographical Information Science **DOI: 10.1080/13658816.2023.2292570** (2023), published December 2023
18. Liu, X., Chen, H., Andris, C.: trajgans: Using generative adversarial networks for geo-privacy protection of trajectory data (vision paper). In: Location Privacy and Security Workshop (LoPaS) (2018)
19. Ma, T., Song, F.: A trajectory privacy protection method based on random sampling differential privacy. ISPRS International Journal of Geo-Information **10**(7), 454 (2021). https://doi.org/10.3390/ijgi10070454
20. Merhi, J., Buchholz, E., Kanhere, S.S.: Synthetic trajectory generation through convolutional neural networks. In: Proceedings of the 21st Annual International Conference on Privacy, Security & Trust (PST 2024). IEEE (2024)
21. Naghizade, E., Kulik, L., Tanin, E., Bailey, J.: Privacy- and context-aware release of trajectory data. ACM Transactions on Spatial Algorithms and Systems **6**(1), 1–25 (2020). https://doi.org/10.1145/3363449
22. Newson, P., Krumm, J.: Hidden markov map matching through noise and sparseness. In: Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS '09). Seattle, WA, USA (2009)
23. Ozeki, R., Yonekura, H., Rizk, H., Yamaguchi, H.: Balancing privacy and utility of spatio-temporal data for taxi-demand prediction. In: 2023 24th IEEE International Conference on Mobile Data Management (MDM). pp. 215–220 (2023). https://doi.org/10.1109/MDM58254.2023.00044
24. Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H.B., Vassilvitskii, S., Chien, S., Thakurta, A.: How to dp-fy ml: A practical guide to machine learning with differential privacy. arXiv preprint arXiv:2303.00654 (2023)
25. Primault, V., Boutet, A., Mokhtar, S.B., Brunie, L.: The long road to computational location privacy: A survey. IEEE Communications Surveys & Tutorials **21**(3), 2772–2793 (2019)
26. Primault, V., Mokhtar, S.B., Lauradoux, C., Brunie, L.: Differentially private location privacy in practice. arXiv preprint arXiv:1410.7744 (2014)
27. Qu, Y., Zhang, J., Li, R., Zhang, X., Zhai, X., Yu, S.: Generative adversarial networks enhanced location privacy in 5g networks. SCIENCE CHINA Information Sciences **63**(220303:1–220303:12) (2020). https://doi.org/10.1007/s11432-019-2834-x
28. Rao, J., Gao, S., Kang, Y., Huang, Q.: Lstm-trajgan: A deep learning approach to trajectory privacy protection. In: Leibniz International Proceedings in Informatics (LIPIcs), GIScience. vol. 177, pp. 1–16 (2021). https://doi.org/10.4230/LIPIcs.GIScience.2021.I.12
29. Rao, J., Gao, S., Zhu, S.: Cats: Conditional adversarial trajectory synthesis for privacy-preserving trajectory data publication using deep learning approaches. International Journal of Geographical Information Science (2023), compiled September 22, 2023
30. Wang, H., Li, Z.: Region representation learning via mobility flow. In: Proceedings of CIKM'17. p. 10 pages. CIKM '17, Association for Computing Machinery, Singapore, Singapore (2017). https://doi.org/10.1145/3132847.3133006

31. Wang, J., Lin, Y., Li, Y.: Gtg: Generalizable trajectory generation model for urban mobility. In: Proceedings of the AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (2025)
32. Wang, L., Chen, M., Zhang, W.: Statistical framework for taxi route anomaly detection using z-score normalization. Transportation Research Part C: Emerging Technologies **115**, 102–118 (2020)
33. Wu, Y., Fang, J., Chen, W., Zhao, P., Zhao, L.: Safety: A spatial and feature mixed outlier detection method for big trajectory data. Information Processing and Management **61**, 103679 (2024)
34. Zhang, L., Mbuya, J., Zhao, L., Pfoser, D., Anastasopoulos, A.: End-to-end trajectory generation - contrasting deep generative models and language models. ACM Transactions on Spatial Algorithms and Systems **2**(ART) (2025). https://doi.org/10.1145/3716892
35. Zhang, Y., Li, F., Wang, H.: ibat: Isolation-based anomaly detection for taxi trajectory data. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1887–1896 (2019)

A Appendix

A.1 Appendix Section

A.2 Appendix Section