



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 Kędzia¹[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. [18] 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. [6] 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. [22]’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 [22], 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. [7]’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. [11]’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. [22] 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. [4] 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. [19] 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 [15] to sophisticated deep learning frameworks [3,17], 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 anomaly detection [22], the 95% re-identification risk that makes real trajectory data unsuitable for research sharing [16], 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. Traditional privacy-preserving mechanisms like k-anonymity and differential privacy create utility-privacy trade-offs that render data unsuitable for complex analytical tasks [9], 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 [3], but success requires solving complex pattern preservation problems across spatial, temporal, and behavioral dimensions [10,14].

Methodological Approaches to Synthetic Generation Synthetic trajectory generation methods broadly fall into two categories: knowledge-driven approaches that rely on mobility theories and empirical patterns, and data-driven approaches that learn trajectory distributions directly from data [10,17].

Knowledge-driven methods, including gravity models and intervening opportunities models, offer interpretability and require minimal training data [13]. However, these approaches often rely on simplified assumptions about human mobility that fail to capture fine-grained behavioral complexities necessary for effective anomaly detection [10]. The Act2Loc framework demonstrates how hybrid approaches can combine machine learning for activity sequence generation with mechanistic models for spatial location selection, addressing some limitations of purely knowledge-driven methods [13].

Data-driven approaches have gained prominence for their ability to capture complex non-linear relationships in trajectory data. Generative Adversarial Networks (GANs) represent the most widely adopted approach, with frameworks like TrajGen successfully generating trajectories that preserve spatial-temporal characteristics [3]. However, GAN-based methods face challenges with training stability and mode collapse, particularly for vehicle trajectories where temporal dependencies prove difficult to capture despite satisfactory spatial modeling [1].

Recent innovations include two-stage generation frameworks that address error accumulation problems in sequential generation. The TS-TrajGen approach first generates structural regions on road networks, then produces continuous trajectories guided by these regional patterns, effectively integrating domain knowledge with model-free learning [8]. This addresses a critical limitation where traditional stochastic generation processes suffer from continuity problems on road networks.

Advanced Architectural Approaches The sequential nature of trajectories has motivated extensive use of Recurrent Neural Networks and Long Short-Term Memory networks. The RMTTP framework demonstrates how RNNs can simultaneously model event timings and spatial markers using recurrent architectures [5]. However, RNN-based GANs often exhibit less stable training compared to CNN models and struggle with convergence issues [14].

Convolutional Neural Networks have shown promise through novel transformation approaches. The Reversible Trajectory-to-CNN Transformation (RTCT) method adapts trajectories into formats suitable for CNN-based models, with Conv1D layers demonstrating superior performance for capturing spatial distributions compared to RNN-based approaches [14]. However, CNN-based methods face challenges in replicating sequential and temporal properties effectively.

Language model-inspired approaches represent an emerging paradigm that conceptualizes trajectories as sequences, with each spatial-temporal point acting as a "word" [20]. Their autoregressive nature enables capture of inherent dependencies, while training on finite vocabulary of locations implicitly enforces spatial-temporal validity constraints [10].

Cross-city generalization has emerged as a critical research direction, with the GTG framework demonstrating how space syntax theory can extract topological features of road networks to learn invariant mobility patterns across different urban environments [17]. This addresses a fundamental limitation where most generation methods require extensive retraining for new geographical contexts.

Privacy Integration and Evaluation The integration of formal privacy guarantees with synthetic generation presents ongoing challenges. The PATE-GAN framework provides differential privacy guarantees by modifying the Private Aggregation of Teacher Ensembles approach for GANs [9]. Alternative approaches include k-anonymity integration through conditional adversarial training on anonymized trajectory matrices [16] and DP-SGD integration with trajectory GANs for stronger privacy guarantees [14].

Evaluation methodologies have evolved from basic similarity metrics to comprehensive utility and privacy assessments. Macro-level metrics like Jensen-Shannon divergence assess overall distribution similarity, while micro-level metrics such as Hausdorff distance evaluate individual trajectory characteristics [10]. Privacy assessment employs metrics like Trajectory-User Linking and Home Location Clustering to measure re-identification difficulty [16]. The Synthetic Ranking Agreement metric evaluates whether relative performance rankings of predictive models are preserved when trained on synthetic versus real data [9].

Implications for Anomaly Detection Research Synthetic trajectory generation directly addresses key requirements for anomaly detection research identified in previous studies. The preservation of origin-destination flow patterns and spatial traversal sequences aligns with needs of isolation-based detection methods like iBAT [22]. For deep learning approaches, synthetic data can provide the large, diverse datasets necessary for training while maintaining the subtle temporal patterns required for autoencoder-based detection [7].

The controlled nature of synthetic datasets enables systematic evaluation of anomaly detection algorithms with known ground truth labels, addressing the parameter sensitivity and labeled data scarcity issues that plague traditional evaluation approaches. This capability is particularly valuable for developing robust anomaly detection systems that can handle the complexity and scale of

modern urban transportation networks while maintaining privacy protection for sensitive location data.

2.3 Privacy Protection Methods

- ▷ *Traditional Privacy Techniques*
- ▷ *Differential privacy – trajectory noise injection [21]*
- ▷ *k-Anonymity methods – spatial cloaking, utility preservation [12]*
- ▷ *Privacy-utility trade-offs – balancing protection and research utility [ADD CITATION]*
- ▷ *Privacy in Synthetic Data*
- ▷ *Synthetic data as privacy solution – avoiding direct exposure [ADD CITATION]*
- ▷ *Attack resistance – membership inference, reconstruction attacks [ADD CITATION]*
- ▷ *Privacy validation methods – measuring protection effectiveness [ADD CITATION]*
- ▷ *Research Gaps and Challenges*
- ▷ *Privacy constraints – limited real data access for research [ADD CITATION]*
- ▷ *Anomaly pattern preservation gap – maintaining detection characteristics [ADD CITATION]*
- ▷ *Comprehensive framework need – integrated privacy-preserving anomaly detection [ADD CITATION]*

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. 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>
4. Chen, J., Liu, X.: Temporal context-aware route anomaly detection in urban transportation. IEEE Transactions on Intelligent Transportation Systems **22**(8), 4892–4903 (2021)
5. 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)
6. He, J., Zhang, P., Liu, G.: Enhanced dbscan with multiple distance metrics for trajectory anomaly detection. Expert Systems with Applications **168**, 114–129 (2020)
7. Huang, Z., Li, J., Chen, R.: Lstm autoencoders with attention mechanisms for trajectory anomaly detection. Neural Networks **142**, 256–271 (2021)
8. 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)
9. 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)
10. 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>
11. 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)
12. Liu, H., Wang, D., Li, X.: Enhanced k-anonymity for trajectory data with improved utility preservation. Information Sciences **598**, 45–62 (2023)
13. 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
14. 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)
15. 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)

16. 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
17. 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)
18. 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)
19. 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)
20. 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>
21. Zhang, M., Liu, B., Chen, F.: Differentially private trajectory synthesis for location privacy protection. *ACM Transactions on Privacy and Security* **26**(2), 1–28 (2023)
22. 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