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

你好，世界！Contribution Title

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

[1] Vrije Universiteit Amsterdam, Amsterdam
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
{abc,lncs}@uni-heidelberg.de

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 transportation systems, particularly taxi services, play a crucial role in the mobility infrastructure of large cities worldwide. These services provide essential connectivity, filling gaps in public transportation networks and offering door-to-door convenience for millions of passengers daily. The efficiency and reliability of taxi operations directly impact urban traffic patterns, economic productivity, and citizen satisfaction with city services.

However, recent research has revealed significant inefficiencies in taxi route selection, with studies consistently showing that drivers often deviate from optimal paths. While some deviations may be justified by real-time traffic conditions or passenger preferences, others stem from more concerning causes including driver inexperience, deliberate route manipulation for fare maximization, or potentially malicious behavior. These routing anomalies not only affect passenger costs and travel times but also contribute to urban congestion and environmental impacts through unnecessary fuel consumption and emissions.

Artificial intelligence technologies, particularly machine learning approaches for anomaly detection, offer promising solutions for identifying and addressing these routing inefficiencies. Various AI methodologies have been developed to detect trajectory anomalies, ranging from classical statistical approaches to modern deep learning techniques. These methods can automatically identify suspicious route patterns that deviate significantly from normal driving behavior, enabling transportation authorities and taxi companies to implement corrective measures.

However, existing approaches face several critical limitations. Most traditional anomaly detection methods struggle with the complexity and contextual nature of urban routing decisions, often producing high false positive rates when applied to real-world taxi data. More sophisticated deep learning approaches, while achieving better accuracy, require extensive labeled datasets and lack the interpretability necessary for practical deployment in regulatory contexts. Furthermore, the sensitive nature of location data raises significant privacy concerns, limiting the availability of real-world datasets for research and deployment.

Privacy-preserving techniques, particularly synthetic data generation, emerge as a promising solution to address these data availability and privacy constraints. By creating artificial datasets that preserve the statistical properties of real trajectory data while protecting individual privacy, synthetic data enables the development and evaluation of anomaly detection systems without compromising passenger confidentiality. However, current synthetic data generation methods for trajectory data remain limited, particularly for capturing the complex spatial-temporal patterns inherent in urban taxi routes.

This study proposes a novel approach for generating synthetic taxi trajectory datasets that preserves the statistical and behavioral properties necessary for effective anomaly detection while addressing critical data privacy concerns. We focus specifically on creating realistic synthetic route data that maintains the complex spatial-temporal patterns inherent in urban taxi operations, enabling privacy-preserving research and development in trajectory anomaly detection systems.

The contribution of this work is threefold: (1) we develop an isolation forest methodology specifically adapted for urban taxi trajectory anomaly detection to establish ground truth patterns, (2) we introduce a comprehensive synthetic trajectory data generation framework that preserves both statistical properties and anomaly characteristics of real taxi routes, and (3) we provide extensive evaluation demonstrating that synthetic data maintains the essential characteristics necessary for effective anomaly detection while providing strong privacy guarantees.

2 Literature Review

This review examines two key research areas relevant to our work: route anomaly detection and synthetic route data generation. We focus specifically on the unique challenges of route data, which differs significantly from other types of data due to its spatial-temporal nature and privacy sensitivity.

2.1 Route Anomaly Detection

Isolation-Based Approaches The seminal work of Zhang et al. [4] established isolation forests as a paradigm-shifting approach for route anomaly detection through their iBat framework. Their method demonstrated remarkable computational efficiency while achieving superior detection accuracy on taxi route data, establishing a foundation for efficient route analysis.

Density-Based Methodologies Building upon earlier clustering paradigms, He et al. [2] advanced density-based anomaly detection by integrating pattern information mining with enhanced DBSCAN clustering. Their incorporation of multiple distance metrics, including DTW and Hausdorff distances, addressed the need for sophisticated similarity measures in route comparison, which is crucial for identifying anomalous paths.

Deep Learning Approaches The emergence of deep learning approaches marked a significant departure from classical methods. Huang et al. [3] pioneered the application of LSTM autoencoders with attention mechanisms, addressing the temporal sequence modeling limitations in route analysis. The recent breakthrough of Li et al. [1] with their DiffTAD framework represents the current pinnacle of generative model applications to route anomaly detection.

2.2 Synthetic Route Data Generation

Unique Challenges of Route Data Route data presents unique challenges for synthetic generation that distinguish it from other types of data. Unlike simple point data or time series, routes contain complex spatial-temporal dependencies, must respect physical constraints (e.g., road networks), and often reveal sensitive information about individuals' movements and habits. These characteristics make traditional synthetic data generation approaches inadequate for route data.

Privacy-Preserving Route Generation The field of synthetic route generation has evolved to address these unique challenges. Early approaches focused on simple perturbation techniques, but these often failed to preserve the complex patterns of real routes. Recent advances in differential privacy and k-anonymity have provided stronger privacy guarantees while attempting to maintain route characteristics, though challenges remain in balancing privacy and utility.

Integration with Anomaly Detection The intersection of synthetic route generation and anomaly detection remains relatively unexplored. While synthetic data has been used to augment training sets for anomaly detection in various domains, the specific challenges of generating realistic anomalous routes while maintaining privacy have received limited attention. This gap in the literature is particularly significant given the sensitivity of route data and the need for privacy-preserving anomaly detection systems.

Synthesis and Future Directions The trajectory of research from classical methods through deep learning approaches reveals a persistent tension between computational efficiency and detection sophistication in route analysis. While deep learning approaches have achieved remarkable performance gains, they have simultaneously highlighted the enduring value of efficient methods for large-scale deployment scenarios. The integration of synthetic route generation with anomaly detection presents an opportunity to address both privacy concerns and data availability limitations, enabling continued research advancement in route analysis.

Our proposed approach addresses the critical need for privacy-preserving route data by developing comprehensive synthetic generation techniques that maintain the statistical and behavioral characteristics essential for anomaly detection research. This approach leverages the efficiency principles established by classical methods while addressing the unique challenges of route data privacy and utility preservation.

3 Methodology

3.1 Isolation Forest for Trajectory Analysis

Algorithm Implementation

Key Adaptations for Trajectory Data

- Feature engineering for route characteristics (distance, duration, deviation metrics)
- Temporal segmentation handling for variable trip lengths
- Geographic normalization across different urban zones

3.2 Statistical Pattern Extraction

Pattern Categories **Spatial Distributions**

Temporal Patterns

Behavioral Characteristics

Anomaly Signatures

3.3 Enhanced Anomaly Detection

Exception Handling Framework **Traffic-Induced Deviations**

Passenger-Requested Deviations

Construction and Event Impacts

Multi-Scale Analysis

- **Micro-level**: [IMPLEMENT] 100-meter segment analysis
- **Meso-level**: [IMPLEMENT] Complete trajectory analysis
- **Macro-level**: [IMPLEMENT] Driver behavior profiling

3.4 Synthetic Trajectory Data Generation

Generation Framework

1. **Pattern Modeling**: [IMPLEMENT] Statistical models for extracted patterns
2. **Route Simulation**: [IMPLEMENT] Probabilistic route generation
3. **Anomaly Injection**: [IMPLEMENT] Systematic anomaly introduction
4. **Noise Addition**: [IMPLEMENT] Realistic GPS error simulation
5. **Validation**: [IMPLEMENT] Quality assurance procedures

Privacy Preservation Mechanisms

- **Statistical Aggregation**: [IMPLEMENT] Aggregate pattern usage only
- **Differential Privacy**: [IMPLEMENT] ε-differential privacy with $\varepsilon =$ [VALUE TO BE DETERMINED]
- **k-Anonymity**: [IMPLEMENT] k = [VALUE TO BE DETERMINED] trajectory indistinguishability

Quality Assurance Framework

- **Distribution tests**: [IMPLEMENT] KS tests, chi-square tests
- **Performance validation**: [IMPLEMENT] Cross-training evaluation
- **Utility assessment**: [IMPLEMENT] Research application validation

4 Data and Preprocessing

4.1 Dataset Description

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

- **Data source**: Beijing taxi GPS tracking devices
- **Geographic coverage**: Beijing metropolitan area
- **Temporal coverage**: 25 November 2019 – 1 December 2019
- **Data volume**: 16GB per day of raw GPS data
- **Licensing**: [TO BE SPECIFIED] - Data agreement and usage conditions

4.2 Data Preprocessing

Data Quality Issues Analysis

- **GPS accuracy variations**: [ANALYZE] Signal loss patterns in urban areas
- **Sampling rate inconsistencies**: [ANALYZE] Time interval variations
- **Missing trajectory segments**: [ANALYZE] Data gap patterns and causes
- **Outlier coordinates**: [ANALYZE] Erroneous GPS coordinate frequency

Preprocessing Pipeline Implementation

1. **Coordinate Validation**: [IMPLEMENT]
2. **Trajectory Segmentation**: [IMPLEMENT]
3. **Gap Interpolation**: [IMPLEMENT]
4. **Map Matching**: [IMPLEMENT]
5. **Feature Extraction**: [IMPLEMENT]

Quality Assessment Results Table 1 will present key statistics before and after preprocessing.

Metric	**Before Preprocessing**	**After Preprocessing**
Total GPS points	0	0
Valid trajectories	0	0
Average trip length (km)	0	0
Average trip duration (min)	0	0
Data completeness (%)	0	0

Table 1. Dataset statistics before and after preprocessing

5 Experimental Setup and Results

5.1 Experimental Design

Evaluation Phases

1. **Anomaly Detection Baseline**: [IMPLEMENT] Ground truth establishment on real data
2. **Synthetic Data Quality Assessment**: [IMPLEMENT] Fidelity and utility evaluation
3. **Privacy Preservation Validation**: [IMPLEMENT] Privacy guarantee assessment

Anomaly Detection Method Comparison

- **Baseline Method**: [IMPLEMENT] Simple statistical thresholds
- **Standard Isolation Forest**: [IMPLEMENT] Traditional approach
- **Enhanced Isolation Forest**: [IMPLEMENT] Our improved approach

5.2 Anomaly Detection Results

Table 2 will present comparative performance on real trajectory data.

Method	**Precision**	**Recall**	**F1-Score**	**Comments**
Baseline	[RESULTS]	[RESULTS]	[RESULTS]	[ANALYSIS]
Standard Isolation Forest	[RESULTS]	[RESULTS]	[RESULTS]	[ANALYSIS]
Enhanced Isolation Forest	[RESULTS]	[RESULTS]	[RESULTS]	[ANALYSIS]

Table 2. Performance comparison of anomaly detection methods on real data

5.3 Synthetic Data Quality Evaluation

Statistical Fidelity Assessment **Distribution Comparisons**

- **Distance distributions**: [IMPLEMENT] Trip length pattern analysis
- **Duration distributions**: [IMPLEMENT] Travel time characteristic comparison
- **Spatial coverage**: [IMPLEMENT] Geographic distribution analysis
- **Temporal patterns**: [IMPLEMENT] Time-of-day and day-of-week pattern comparison
- **Speed profiles**: [IMPLEMENT] Velocity and acceleration pattern analysis

Statistical Test Results
Table 3 will show comparison between real and synthetic datasets.

Metric	Real Data	Synthetic Data	Difference (%)
Avg. trip distance (km)	0	0	0
Avg. trip duration (min)	0	0	0
Spatial coverage (km^2)	0	0	0
Peak hour ratio	0	0	0
Anomaly rate (%)	0	0	0

Table 3. Statistical comparison between real and synthetic trajectory datasets

Anomaly Preservation Evaluation **Cross-Training Experiments**

IMPLEMENT Train models on synthetic data, test on real data
IMPLEMENT Train models on real data, test on synthetic data
IMPLEMENT Compare performance across training scenarios
IMPLEMENT Validate anomaly characteristic preservation

Detection Challenge Preservation

Utility Validation

- **Algorithm Development**: [TEST] New method development on synthetic data
- **Parameter Optimization**: [TEST] Hyperparameter tuning transferability
- **Research Reproducibility**: [VALIDATE] Replication capability for other researchers

5.4 Privacy Preservation Assessment

Attack Resistance Testing **Membership Inference Attacks**

Trajectory Reconstruction Attacks

Location Privacy Protection

Privacy-Utility Trade-off Analysis

5.5 Computational Performance Analysis

Scalability Analysis

- **Pattern Extraction**: [MEASURE] Processing time vs. dataset size
- **Synthetic Generation**: [MEASURE] Generation time vs. output size
- **Privacy Mechanisms**: [MEASURE] Privacy overhead analysis

Resource Requirements

6 Conclusion and Future Work

6.1 Research Contributions Summary

Primary Contributions

1. **Synthetic Trajectory Data Generation Framework**: [SUMMARIZE] Development of comprehensive privacy-preserving synthetic data generation methodology
2. **Enhanced Isolation Forest for Trajectory Analysis**: [SUMMARIZE] Adaptation of isolation forests for urban taxi trajectory anomaly detection
3. **Privacy-Utility Trade-off Analysis**: [SUMMARIZE] Comprehensive evaluation of privacy preservation vs. data utility

6.2 Research Impact and Applications

Academic Impact

DISCUSS Advancement of privacy-preserving data generation techniques

DISCUSS Contribution to trajectory anomaly detection methodologies

DISCUSS Framework for evaluating synthetic data quality

Practical Applications

DISCUSS Transportation authority anomaly detection systems

DISCUSS Privacy-compliant research data sharing

DISCUSS Taxi company route optimization and monitoring

6.3 Limitations and Challenges

Current Limitations

IDENTIFY Computational complexity limitations

IDENTIFY Data dependency requirements

IDENTIFY Privacy-utility trade-off constraints

Technical Challenges

DISCUSS Scalability challenges with large datasets

DISCUSS Parameter tuning complexity

DISCUSS Validation methodology limitations

6.4 Future Research Directions

Methodological Extensions

- PROPOSE: Advanced privacy mechanisms (federated learning, homomorphic encryption)
- PROPOSE: Deep learning integration for pattern modeling
- PROPOSE: Multi-modal data integration (traffic, weather, events)

Evaluation Framework Extensions

- PROPOSE: Longitudinal evaluation over extended time periods
- PROPOSE: Cross-city validation and generalization testing
- PROPOSE: User study integration for practical utility assessment

6.5 Concluding Remarks

References

1. Li, C., Feng, G., Li, Y., Liu, R., Miao, Q., Chang, L.: Difftad: Denoising diffusion probabilistic models for vehicle trajectory anomaly detection. Knowledge-Based Systems **286**, 111387 (2024). https://doi.org/https://doi.org/10.1016/j.knosys.2024.111387, https://www.sciencedirect.com/science/article/pii/S0950705124000224
2. Ming, H., Yuting, C., Qiang, L., Bo, Z., Gongda, Q.: . (6), 49–54 (2019)
3. Shi-chen, H., Chun-fu, S., Juan, L., Zong-tao, D.: . **21**(3), 47–54 (2021)
4. Zhang, D., Li, N., Zheng, Y., Ramamohanarao, K., Zhao, Z.: ibat: Detecting anomalous taxi trajectories from gps traces. In: UbiComp'11: Proceedings of the 13th international conference on Ubiquitous computing. pp. 99–108. ACM (2011). https://doi.org/10.1145/2060091.2060106

[Fix Chinese chars not displaying]