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 spatio-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 spatio-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 demonstrate the essential properties that synthetic trajectory data must preserve to maintain utility for anomaly detection research. Different detection methods rely on fundamentally distinct trajectory characteristics, establishing specific requirements for data generation.

Distance-based methods like Wang et al. [33] 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. This method establishes two critical requirements for synthetic data generation: preserving origin-destination flow patterns and maintaining consistent spatial traversal sequences between locations.

Traditional methods also demonstrate a significant research gap that synthetic data generation addresses. Most approaches struggle with parameter sensitivity and insufficient labeled anomaly data [35], creating challenges for systematic evaluation of new detection algorithms. Synthetic generation provides a solution through controlled datasets with known anomaly labels and adjustable parameters for systematic evaluation.

Deep Learning Approaches Deep learning approaches present distinctive challenges for synthetic data generation, as these methods depend on learning complex patterns that traditional approaches cannot capture.

Autoencoder-based detection, exemplified by Huang et al. [11]'s LSTM-AE-Attention model, operates by learning to reconstruct normal trajectory patterns. Anomalous trajectories that exhibit poor reconstruction quality are identified as suspicious. This approach establishes a critical requirement for synthetic data: preservation of subtle temporal patterns and sequence dependencies that characterize real trajectories, as their absence would compromise reconstruction-based detection effectiveness. The study also identifies a practical challenge where real datasets exhibit significant imbalance, with approximately 12 normal trajectories for every anomalous one, complicating training processes.

More recent work with diffusion models, such as Li et al. [16]'s DiffTAD, demonstrates that synthetic trajectory generation can be directly applied for anomaly detection. Their approach treats trajectory generation as a denoising process, achieving significantly superior performance compared to traditional methods. This suggests potential adaptation of synthetic data generation techniques developed for privacy protection to anomaly detection applications.

Deep learning methods present specific requirements for synthetic data research, requiring large training datasets and performing optimally when learning from diverse trajectory patterns. Synthetic data generation addresses these requirements by providing abundant, diverse trajectory data that preserves essential characteristics necessary for effective anomaly detection.

Spatio-Temporal Pattern Analysis Identifying the critical patterns in trajectory data defines the preservation requirements for synthetic generation. Research demonstrates that trajectories exhibit multi-level structural properties essential for anomaly detection algorithm effectiveness.

At the spatial level, Zhang et al. [35] demonstrate that converting continuous GPS traces into grid-based symbolic sequences achieves effective anomaly detection. This indicates that synthetic data need not precisely replicate individual GPS coordinates, but must preserve the sequence of spatial regions traversed by vehicles. Their approach effectively manages variable GPS sampling rates, which is significant given that synthetic data may exhibit different temporal characteristics than real data.

Temporal patterns exhibit greater complexity than initially apparent. Chen et al. [6] demonstrate that normal behavior definitions vary significantly based on temporal context - routes considered normal during off-peak hours may appear highly suspicious during rush hour periods. This requires synthetic data generation to preserve time-dependent behavioral patterns in addition to spatial accuracy.

Large-scale analysis provides significant insights, as demonstrated by Balan et al. [2]'s study of 250 million taxi trips. Their findings indicate that urban mobility follows predictable patterns, with normal routes clustering around preferred paths between locations, and these patterns exhibit sufficient repetition to enable statistical prediction. For synthetic data generation, this emphasizes the importance of preserving origin-destination flow patterns and route clustering rather than generating entirely novel trajectory types.

Scalability represents an important practical consideration for synthetic data generation. Wu et al. [34] demonstrate that modern anomaly detection requires distributed processing approaches to handle large datasets effectively. Consequently, synthetic data generation methods must produce datasets of sufficient scale and appropriate structure 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 [5,32]. This evolution is 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.

Three critical problems drive this development. First, the parameter sensitivity and labeled data scarcity issues identified in trajectory anomaly detection

research (Section 2.1) [35] make it difficult to systematically evaluate detection algorithms. Second, the high re-identification risk that makes real trajectory data unsuitable for research sharing [29] creates fundamental data access barriers. Third, the need for reproducible evaluation frameworks that traditional privacy methods cannot provide limits research reproducibility.

This convergence shows 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 [14]. Meanwhile, 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]. However, 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 are essential for practical deployment of synthetic trajectory generation systems.

Evolution of Generation Approaches The development of synthetic trajectory generation shows 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. These insights directly address the spatial traversal sequence requirements identified in isolation-based detection methods like iBAT (Section 2.1).

The deep learning transition created a paradigm shift through GAN-based approaches like TrajGen [5]. These approaches demonstrated that neural networks can capture complex spatio-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. This directly impacts the subtle temporal patterns and sequence dependencies that autoencoder-based detection methods require (Section 2.1) [11].

These limitations drove architectural innovations including CNN-based transformations [20] for spatial distribution capture and RNN approaches [8] for sequential dependencies. Each approach addresses 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. These approaches integrate domain knowledge with model-free learning. Cross-city generalization research [32] demonstrates scalable pattern extraction across urban environments, directly addressing the distributed processing and scalability requirements identified for modern anomaly detection systems (Section 2.1).

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 (Section 2.1). The RMTPP framework [8] shows how RNNs can model event timings and spatial markers simultaneously. This demonstrates the importance of temporal pattern preservation for sequence-dependent detection methods like LSTM-AE-Attention models.

However, RNN-based GANs exhibit training instability compared to CNN models [20]. This creates trade-offs between temporal modeling capability and training reliability. Convolutional approaches like the RTCT method [20] solve spatial distribution challenges through novel data transformations. Conv1D layers demonstrate superior performance for capturing spatial distributions needed for density-based anomaly detection methods that rely on local neighborhood structures (Section 2.1).

This research shows 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 spatio-temporal point acts as a "word" [36]. This approach addresses both sequential dependencies and spatial constraints simultaneously through autoregressive modeling. Training on finite vocabulary of locations implicitly enforces spatio-temporal validity constraints [15].

This paradigm shift leverages broader AI advances to potentially resolve the architectural trade-offs identified in earlier approaches. The approach maintains compatibility with privacy protection mechanisms and cross-city generalization through space syntax theory [32].

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 privacy-preserving integration [29] constrains input representations and DP-SGD integration [20] constrains optimization processes. These requirements demonstrate 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 high re-identification risks 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.

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 require the specific origin-destination flow patterns and spatial traversal sequences identified in Section 2.1 [35], 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]. However, 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 the parameter sensitivity and labeled data scarcity issues identified in anomaly detection research (Section 2.1) [35], but current generation methods do not provide systematic evaluation capabilities needed. Cross-city generalization research [32] 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 (Section 2.2), privacy protection (Section 2.3), 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. Such frameworks require seamless integration of the anomaly detection requirements (Section 2.1), generation capabilities, and privacy protection mechanisms examined across this literature review.

2.3 Privacy Protection Methods

Privacy Challenges in Trajectory Data 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]. The inherent challenge for anomaly detection research is balancing privacy protection with the need to preserve complex spatio-temporal patterns that detection algorithms require [4,3,25,21].

The central goal is enabling high-utility trajectory data release without revealing private information about individuals [4,28,18,13,19,21]. This would allow researchers to develop and test anomaly detection systems without requiring direct access to sensitive real-world data [4,28,18].

Traditional Privacy-Preserving Methods and Limitations Traditional privacy-preserving methods like k-anonymity have been 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]. Similarly, conventional approaches such as suppression and generalization techniques struggle with the inherent complexity of trajectory data. These methods often destroy the spatio-temporal relationships that anomaly detection algorithms require. 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].

To achieve robust privacy for synthetic trajectory data, researchers must carefully select the Unit of Privacy (UoP) [4,25]. Protecting individual locations (location-level privacy) in a trajectory is considered the weakest level. This approach 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, offers a more promising balance for deep learning applications [4]. These limitations of traditional methods have driven the development of synthetic data generation approaches as more viable alternatives.

Synthetic Data Generation for Privacy Protection Given these challenges, synthetic trajectory data generation represents a promising alternative to directly protecting original data [4,28,18]. The approach creates new, non-real trajectories that mimic the statistical and behavioral properties of the authentic data. These synthetic trajectories can then be freely shared for research and development without privacy concerns attached to specific individuals [28,18,27].

Deep learning approaches, particularly Generative Adversarial Networks (GANs), have emerged as a key direction for privacy-preserving synthetic trajectory data generation [4,18,28,27]. Liu et al. [18] proposed trajGANs to generate synthetic trajectories that preserve the summary properties of real data and achieve close-to-real-data performance in analysis tasks. These privacy-focused approaches build on the generation capabilities and architectural trade-offs discussed in Section 2.2 to specifically address data protection requirements [28,27,4,24].

For urban taxi operations, generative models have been evaluated on real-world datasets like the T-Drive dataset (Beijing taxi trajectories) and the San Francisco cabs dataset [19,26,25]. These datasets capture the specific spatio-temporal continuity and regularity of taxi movements, which models like LSTM-TrajGAN are designed to preserve [28,18,13]. Preserving spatial and temporal characteristics is crucial for anomaly detection, as anomalies are deviations from learned normal patterns. This requirement creates particular challenges for privacy-preserving methods, as the complex patterns needed for effective anomaly detection (Section 2.1) must be maintained while protecting individual privacy [28,21].

Alternative approaches include DP-driven synthetic methods such as DPT (Differentially Private Trajectory Synthesis), which adapts the Laplacian mechanism and uses hierarchical reference systems to model trajectories [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 while providing formal privacy guarantees [13,27].

Privacy Evaluation and Open Challenges 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. Such an approach would overcome privacy concerns and regulatory hurdles associated with using actual mobility traces [4,28,18]. Synthetic data must support diverse analytical tasks including spatial and temporal analyses, classification, clustering, and anomaly detection while maintaining utility for research purposes [28,7]. The CC-Net system demonstrates privacy-preserved taxi demand prediction, achieving high accuracy while ensuring privacy by design [23].

Despite progress, significant challenges remain unresolved. 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]. The assessment of privacy guarantees requires diligent verification for any proposed synthetic data generation method [4]. Future research must focus on developing 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]. 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].

2.4 Synthesis and Research Framework

The comprehensive examination of trajectory anomaly detection, synthetic data generation, and privacy protection reveals a critical convergence point that defines the research opportunity addressed in this thesis. The three research areas exhibit complementary strengths and limitations that, when properly integrated, create a pathway to address fundamental challenges in privacy-preserving trajectory research.

Convergence of Research Requirements The analysis demonstrates that trajectory anomaly detection, synthetic data generation, and privacy protection share fundamental requirements that must be addressed simultaneously rather than sequentially. Anomaly detection algorithms require specific pattern preservation capabilities: origin-destination flow patterns and spatial traversal sequences for isolation-based methods (Section 2.1), subtle temporal patterns and sequence dependencies for deep learning approaches, and time-dependent behavioral patterns for comprehensive spatio-temporal analysis.

Synthetic data generation approaches have developed sophisticated capabilities to address these requirements through architectural innovations including CNN-based spatial modeling, RNN-based temporal processing, and language model paradigms that handle both spatial and temporal constraints (Section 2.2). However, these generation capabilities face fundamental constraints when privacy protection mechanisms are integrated, as differential privacy guarantees, trajectory-level protection, and attack resistance requirements fundamentally alter training processes and pattern preservation capabilities.

Privacy protection research identifies the critical challenge that traditional privacy-preserving methods destroy the spatio-temporal relationships essential

for accurate anomaly detection (Section 2.3). This creates a design space where privacy constraints and research utility requirements must be jointly optimized. The 95% individual identification risk from just four spatio-temporal points demonstrates why privacy cannot be treated as a post-processing step, but must be integrated throughout the entire research pipeline.

Integrated Framework Requirements The convergence analysis reveals that effective privacy-preserving trajectory anomaly detection requires an integrated framework that addresses five core challenges simultaneously:

Pattern Preservation Under Privacy Constraints. The framework must preserve the specific trajectory characteristics that anomaly detection algorithms require while providing strong privacy guarantees. This requires understanding how privacy protection mechanisms affect the spatial traversal sequences, temporal dependencies, and behavioral patterns that different detection methods depend on.

Scalable Synthetic Generation. The framework must generate synthetic datasets of sufficient scale and diversity to support systematic evaluation of anomaly detection algorithms while maintaining computational efficiency for practical deployment. This addresses the parameter sensitivity and labeled data scarcity issues identified in anomaly detection research.

Comprehensive Privacy Protection. The framework must provide robust privacy protection against sophisticated attacks while preserving utility for anomaly detection research. This requires careful selection of privacy units, integration of multiple protection mechanisms, and evaluation against both membership inference and reconstruction attacks.

Systematic Evaluation Capabilities. The framework must enable systematic evaluation of anomaly detection methods through controlled synthetic datasets with known anomaly characteristics. This addresses the reproducibility and comparison challenges that limit current anomaly detection research.

Practical Deployment Considerations. The framework must address the computational requirements, scalability constraints, and cross-city generalization needs for practical deployment in urban transportation systems.

Research Contribution and Methodology Framework This thesis addresses the identified convergence point by developing a comprehensive framework that integrates isolation forest-based anomaly detection, statistical pattern extraction, and privacy-preserving synthetic generation. The approach leverages the strengths identified in each research area while addressing their individual limitations through systematic integration.

The methodology framework builds on isolation forest analysis to understand both normal and anomalous trajectory patterns in real data, extracting the specific statistical and behavioral properties that must be preserved in synthetic generation. The framework implements multiple privacy protection mechanisms designed to work together rather than independently, ensuring that privacy guar-

antees do not compromise the pattern preservation essential for anomaly detection research.

The synthetic generation component addresses the architectural trade-offs identified in existing approaches by combining spatial modeling capabilities with temporal pattern preservation, while the privacy protection mechanisms ensure that the resulting synthetic data provides strong privacy guarantees without destroying research utility. The comprehensive evaluation framework enables systematic assessment of both privacy protection and anomaly detection performance, addressing the evaluation gaps identified across all three research areas.

This integrated approach creates a research contribution that extends beyond individual advances in any single area, providing a complete solution for privacy-preserving trajectory anomaly detection research that addresses the fundamental challenges identified in each research domain while enabling practical deployment in urban transportation systems.

3 Methodology

This research introduces an iterative framework for generating controllable and diverse synthetic anomalies. The methodology bootstraps anomaly generation without requiring pre-labeled anomaly data, instead relying on an unsupervised detector and rule-based curation to iteratively refine the generation model. The process consists of three main phases designed to enrich the synthetic dataset with specific, interpretable anomalies, as formalized in Algorithms 1 and 2.

Algorithm 1 Iterative Anomaly Generation Framework

Require: D_{real}: Dataset of real normal trajectories. N: Number of refinement iterations. M: Number of trajectories to generate per iteration. *Detector*: Unsupervised anomaly detection model. *Rules*: Heuristic rules for anomaly categorization.

Ensure: G_N: Final generative model for conditional anomaly generation. $D_{anomalous}$: Labeled dataset of synthetic anomalies.

Initialize and train generative model G_0 on D_{real}.
$D_{enriched} \leftarrow D_{real}$.
$D_{anomalous} \leftarrow \emptyset$.
for $i \leftarrow 1$ to N **do**
    $D_{new\_anomalies} \leftarrow \text{MineAnomalies}(G_{i-1}, M, Detector, Rules)$. ▷ See Algorithm 2
    $D_{anomalous} \leftarrow D_{anomalous} \cup D_{new\_anomalies}$.
    $D_{enriched} \leftarrow D_{enriched} \cup D_{new\_anomalies}$.
    Initialize and train new model G_i on $D_{enriched}$.
end for
return $G_N, D_{anomalous}$.

Algorithm 2 Unsupervised Anomaly Mining and Curation ('MineAnomalies')

Require: G_{in}: Input generative model. M: Number of trajectories to generate. *Detector*: Unsupervised anomaly detection model. *Rules*: Heuristic rules for anomaly categorization.

Ensure: $D_{new\_anomalies}$: Labeled dataset of new synthetic anomalies.

$D_{synthetic} \leftarrow$ Generate M trajectories using G_{in}.
Train *Detector* on $D_{synthetic}$.
$D_{potential} \leftarrow$ Identify outliers in $D_{synthetic}$ using *Detector*.
$D_{new\_anomalies} \leftarrow \emptyset$.
for each $traj$ in $D_{potential}$ **do**
    $label \leftarrow$ Categorize($traj, Rules$).
    if $label \neq$ None **then**
        Add $(traj, label)$ to $D_{new\_anomalies}$.
    end if
end for
return $D_{new\_anomalies}$.

Table 1. Nomenclature for Algorithms 1 and 2

Symbol	Description
D_{real}	A dataset containing real-world, normal trajectories.
$D_{synthetic}$	A dataset of trajectories generated by the model.
$D_{potential}$	A subset of synthetic trajectories identified as potential anomalies.
$D_{anomalous}$	A curated dataset of synthetic trajectories labeled as anomalies.
$D_{enriched}$	The training dataset, augmented with newly curated anomalies.
$D_{new\_anomalies}$	A set of newly discovered and labeled anomalies from an iteration.
G_i	The generative model at iteration i.
G_{in}	The input generative model for the anomaly mining process.
N	The total number of iterative refinement cycles to perform.
M	The number of synthetic trajectories to generate in each iteration.
Detector	The unsupervised machine learning model used to identify outliers.
Rules	A set of heuristics used to categorize potential anomalies.
$traj$	A single trajectory data structure.
label	The specific anomaly category assigned to a trajectory.

3.1 Phase 1: Baseline Synthetic Data Generation

The initial phase focuses on creating a high-fidelity synthetic dataset of normal trajectories, which serves as the foundation for the anomaly mining process.

Core Generation Model (DiffTraj) The DiffTraj architecture—a 1D-CNN-based residual network with attention—is adopted for its demonstrated training stability and ability to generate high-quality, realistic trajectories.

Training on Normal Data The model is trained solely on a dataset of real, normal taxi trajectories to capture the underlying patterns of typical urban mobility.

Output Dataset This phase produces a synthetic trajectory dataset, `synthetic_normal`, which statistically resembles the real normal data but contains no direct copies, thereby ensuring privacy by design.

3.2 Phase 2: Unsupervised Anomaly Mining and Curation

This phase uses an unsupervised approach to discover and categorize anomalous patterns within the baseline synthetic data.

Unsupervised Anomaly Detection A lightweight anomaly detection algorithm (such as Isolation Forest, Autoencoder, or One-Class SVM) is applied to the `synthetic_normal` dataset to identify outliers without requiring pre-existing labels.

Rule-Based Curation and Categorization The detected `potential_anomalies` are filtered and classified into interpretable categories (e.g., speeding, off-road, unusual stops) using heuristic rules based on kinematic and spatial properties.

Labeled Anomaly Subsets This curation process produces labeled anomaly datasets (e.g., `anomalies_speeding`, `anomalies_off_road`) that are ready for model retraining.

3.3 Phase 3: Iterative Refinement and Conditional Generation

The final phase closes the loop by retraining the diffusion model on the enriched dataset to learn and amplify the discovered anomalous patterns, enabling controlled generation.

Enriched Data Retraining The DiffTraj model is retrained on a dataset that combines the original `synthetic_normal` trajectories with the curated, labeled anomalies. The proportion of anomalies is carefully balanced (e.g., 5–10

Iterative Amplification This process—generation, detection, curation, and retraining—can be repeated multiple times. Each iteration further amplifies and diversifies the model's ability to generate complex and varied anomalies.

Controlled Anomaly Generation The refined model supports conditional sampling (e.g., `difftraj.sample(condition="speeding")`), enabling the targeted generation of specific, high-quality anomalies for downstream tasks.

3.4 Model Selection and Justification

The selection of the DiffTraj model and the iterative, bootstrapping methodology is grounded in a pragmatic assessment of technical advantages and research objectives.

Generative Model Choice DiffTraj was selected over alternatives like GANs and VAEs for its superior training stability, high-fidelity sample generation, and inherent privacy-preserving design, which avoids direct replication of real data.

Rationale for Iterative Approach The proposed iterative framework was chosen for its key strategic benefits:
- **No Labeled Data Required:** It bootstraps the anomaly generation process without needing a pre-labeled anomalous dataset, addressing a common bottleneck in anomaly detection research.
- **High Control and Interpretability:** The rule-based curation ensures that generated anomalies correspond to clear, interpretable, and controllable categories.
- **Scalability and Efficiency:** The approach is computationally less intensive than complex latent space manipulation techniques and allows the anomaly dataset to be iteratively expanded and refined.

Trade-off Analysis This methodology prioritizes interpretability, control, and implementation feasibility. It serves as a robust foundation, acknowledging that more complex generative techniques like latent space adversarial training are deferred to future work.

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 Raw Data Statistics and Characteristics

- ▷ *Data Volume Analysis – Daily data sizes, total trajectory counts*
- ▷ *Basic Statistical Properties – Trip durations, distances, spatial coverage*
- ▷ *Temporal Distribution – Time-of-day patterns, weekly variations*
- ▷ *Spatial Coverage – Geographic extent, density distributions*

4.3 Data Source and Privacy Issues

- ▷ *Data Source Documentation – Origin, collection methodology, licensing*
- ▷ *Privacy Risk Assessment – Individual identification risks, sensitive locations*
- ▷ *Regulatory Compliance – Legal requirements and constraints*
- ▷ *Pseudo-anonymisation Strategy – Approach to privacy protection*

4.4 Data Quality Assessment

- ▷ *Data Quality Issues Analysis – Missing data, GPS accuracy, temporal gaps*
- ▷ *Anomalous Data Detection – Identification of erroneous trajectories*
- ▷ *Exclusion Criteria – Examples of excluded data with justifications*
- ▷ *Data Completeness Analysis – Coverage and representativeness evaluation*

4.5 Preprocessing Pipeline

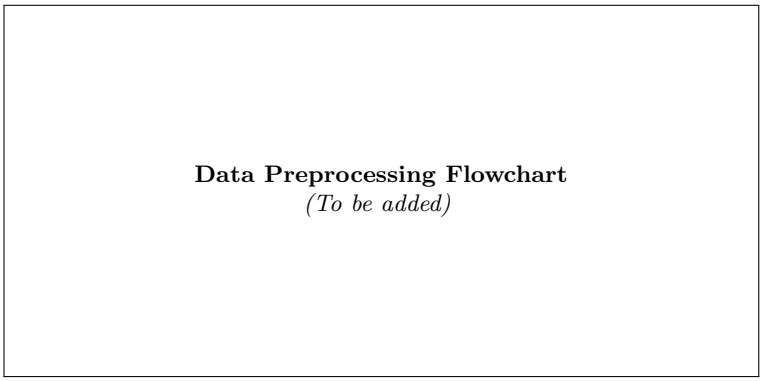

Fig. 1. Data preprocessing pipeline flowchart showing the complete workflow from raw GPS data to LibCity-standardized trajectory datasets.

- ▷ *Data Cleaning Framework – Error correction and outlier removal*
- ▷ *Trajectory Reconstruction – Handling missing GPS points and temporal gaps*
- ▷ *LibCity Format Standardization – Converting data to LibCity's standardized format for compatibility*
  To ensure standardization and reproducibility, this research leverages the LibCity framework [31], which provides standardized data formats and comprehensive tools for spatio-temporal data analysis. The LibCity framework enables seamless integration with multiple datasets and facilitates the implementation of classification and data generation extensions while maintaining compatibility with existing prediction tasks and evaluation frameworks.
- ▷ *Parameter Selection – Preprocessing parameters with justifications*
- ▷ *Quality Control Measures – Validation of preprocessing results*

4.6 Processed Data Statistics

- ▷ *Post-Processing Statistics – Comparison with raw data characteristics*
- ▷ *Quality Improvements – Quantitative assessment of preprocessing impact*
- ▷ *Data Reduction Analysis – Volume and coverage after preprocessing*
- ▷ *Validation Results – Verification of data quality and integrity*

5 Experimental Setup and Results

The evaluation framework is designed to be comprehensive and multi-faceted, assessing the generated data from three critical perspectives: the performance of the anomaly detection system, the quality of the synthetic data, and the robustness of the privacy-preserving mechanisms.

5.1 Experimental Design and Validation Strategy

The experimental design incorporates a cross-city and cross-dataset validation strategy to ensure the generalizability and robustness of the findings.

Multi-Dataset Approach The methodology is validated across multiple datasets, including the public Beijing T-Drive dataset, additional datasets from Chengdu and Xi'an, and a private Beijing dataset. This approach allows for robust testing across diverse urban environments.

Cross-City Generalization A key component of the validation is testing the transferability of learned anomaly patterns between cities (e.g., training a model on Beijing data and testing its performance on Chengdu data) to assess the model's generalization capabilities.

Public vs. Private Dataset Comparison The framework includes a novel comparative analysis between public and private datasets from the same city (Beijing) to investigate potential discrepancies and biases in publicly available data.

5.2 Anomaly Detection Performance

The performance of the anomaly detection system is evaluated using metrics appropriate for imbalanced datasets, where anomalies are rare.

Key Performance Metrics Evaluation focuses on Precision, Recall, and the F1-Score, which provide a balanced view of the detector's ability to correctly identify rare anomalous instances.

AUC-ROC and AUC-PR The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the Area Under the Precision-Recall Curve (AUC-PR) are used to assess the model's overall discriminative power, with AUC-PR being particularly informative for imbalanced class distributions.

5.3 Synthetic Data Quality Evaluation

The quality of the generated synthetic data is assessed using a standardized framework to ensure it is both realistic and useful for downstream tasks.

Standardized Quality Assessment The **SDMetrics** library is employed to systematically evaluate the synthetic data. This includes assessing statistical resemblance to real data (Resemblance), utility for machine learning tasks (Utility), and protection against disclosure (Privacy).

Statistical Fidelity Distribution comparison tests (e.g., Kolmogorov-Smirnov, Jensen-Shannon divergence) are used to quantitatively measure the statistical similarity between the real and synthetic trajectory data distributions for key properties like trip duration and distance.

Downstream Task Performance The utility of the synthetic data is further validated by evaluating the performance of downstream models (e.g., travel time estimation, destination prediction) trained on the synthetic data versus models trained on real data, leveraging the LibCity framework's benchmark tasks.

5.4 Privacy Preservation Assessment

The privacy guarantees of the synthetic data are evaluated through a series of attack simulations designed to test its resilience against re-identification.

Membership Inference Attacks Tests are conducted to determine whether an adversary can successfully identify whether a specific, real trajectory was part of the original training dataset used to create the synthetic data.

Trajectory Reconstruction Attacks The framework is evaluated on its ability to prevent an adversary from reconstructing individual, real-world trajectories from the synthetic dataset.

Privacy-Utility Trade-off A quantitative analysis is performed to measure the balance between the level of privacy protection achieved and the resulting utility of the data for anomaly detection research.

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*
- ▷ *Framework Extension – Classification and data generation extensions to LibCity framework*
- ▷ *Standardization Benefits – Enabling cross-dataset compatibility and reproducible research*

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)
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), http://arxiv.org/abs/1112.2020
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)
18. Liu, X., Chen, H., Andris, C.: trajGANs: Using generative adversarial networks for geo-privacy protection of trajectory data (Vision paper). Location Privacy and Security Workshop (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. LIPIcs, Volume 177, GIScience 2021 **177**, 12:1–12:17 (2020). https://doi.org/10.4230/LIPICS.GISCIENCE.2021.I.12, https://drops.dagstuhl.de/entities/document/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)
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., Jiang, J., Jiang, W., Han, C., Zhao, W.X.: Towards Efficient and Comprehensive Urban Spatial-Temporal Prediction: A Unified Library and Performance Benchmark. arXiv preprint arXiv:2304.14343 (2023)
32. 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)
33. 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)
34. 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)
35. Zhang, D., Li, N., Zhou, Z.H., Chen, C., Sun, L., Li, S.: iBAT: Detecting anomalous taxi trajectories from GPS traces. In: Proceedings of the 13th International Conference on Ubiquitous Computing. pp. 99–108. ACM, Beijing China (Sep 2011). https://doi.org/10.1145/2030112.2030127, https://dl.acm.org/doi/10.1145/2030112.2030127
36. 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

A Appendix

A.1 Appendix Section

A.2 Appendix Section